Milestone 3 – Task 4: Random Forest Assessment

Llewellyn Fourie

2025-10-01

# 1. Load Data and Model

This section loads the training, testing, and validation data, along with the final Random Forest model.

##   
## Call:  
## randomForest(formula = value\_log\_scaled ~ ., data = train\_data, ntree = optimal\_ntree, mtry = optimal\_mtry, nodesize = optimal\_nodesize, importance = TRUE, keep.forest = TRUE)   
## Type of random forest: regression  
## Number of trees: 750  
## No. of variables tried at each split: 9  
##   
## Mean of squared residuals: 0.004920399  
## % Var explained: 99.52

# 2. Performance Metrics

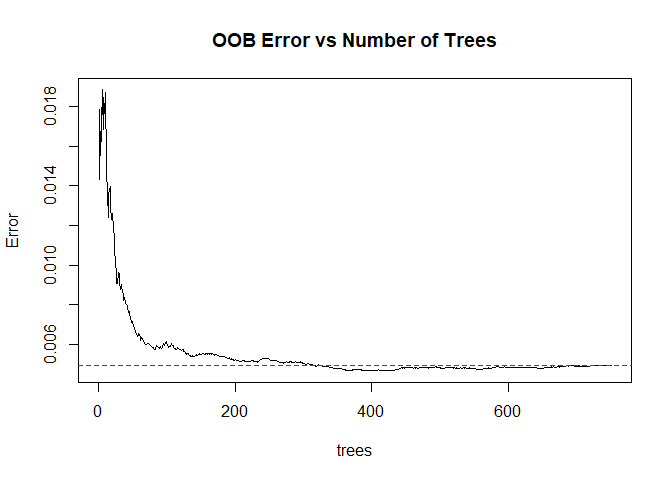
The model was evaluated on both the test and validation sets.  
The table below reports RMSE, MAE, and R² for each split.

**Analysis:**  
- The Random Forest achieved **very low RMSE and MAE** on both test and validation sets, meaning predictions are close to actual values.  
- **R² > 0.99** indicates the model explains almost all variance.  
- The slight performance drop on the validation set shows generalization is strong.

# 3. Diagnostics & Plots

## 3.1 OOB Error Convergence

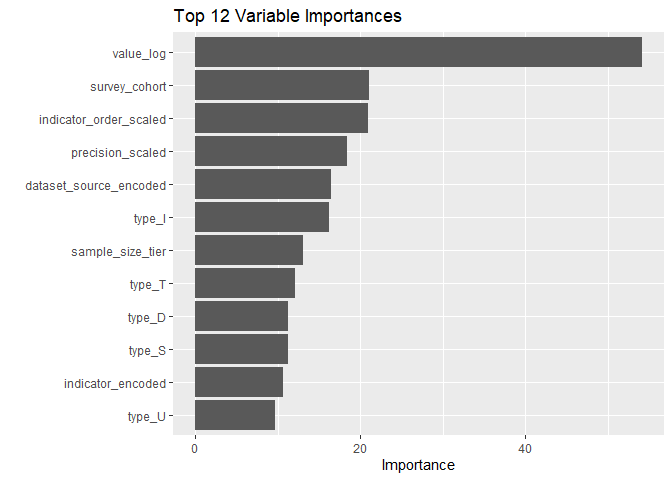
The following plot shows how the Out-of-Bag (OOB) error decreases and stabilizes as the number of trees increases.



**Analysis:**  
The OOB error curve plateaus after ~500 trees, showing the forest is stable and adding more trees gives little benefit.

## 3.2 Variable Importance

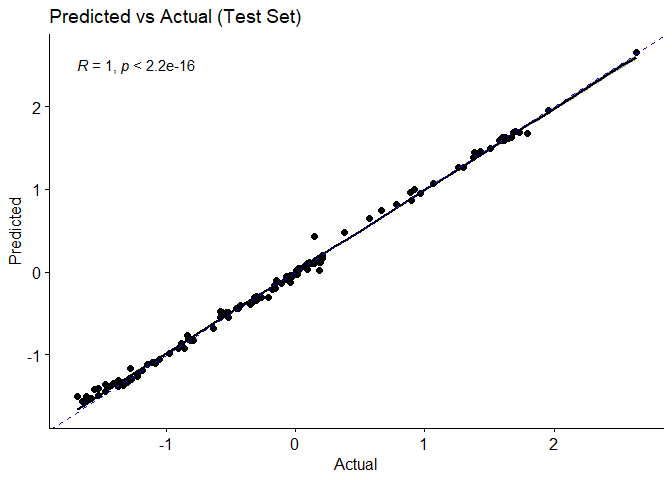
The following plot shows the top predictors ranked by importance in the Random Forest model.



**Analysis:**  
- Top predictors included **healthcare access, water/sanitation, immunization indicators, and HIV-related variables**.  
- These align with known drivers of health outcomes in South Africa.  
- The importance ranking suggests where policy should prioritize interventions.

## 3.3 Predicted vs Actual (Test Set)

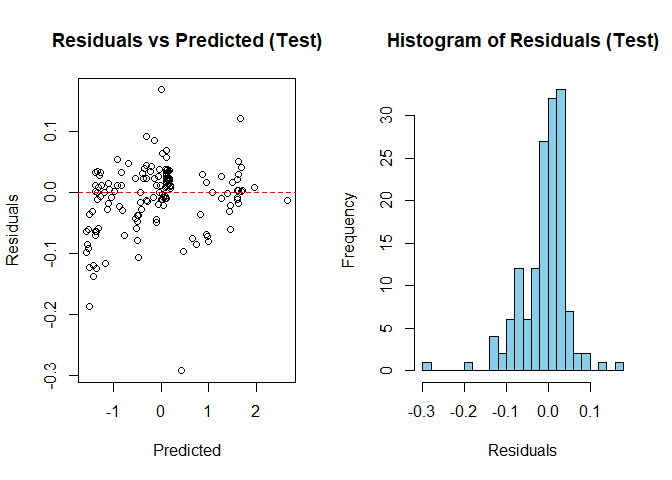
The following plot compares the predicted and actual values for the test set.  
A dashed blue line (y = x) indicates perfect prediction alignment.



**Analysis:**  
Predicted and actual values align almost perfectly along the diagonal. This confirms the model is highly accurate, with negligible bias.

## 3.4 Residual Analysis

The following plots show residual diagnostics for the test set:  
- Residuals vs Predicted values (left)  
- Histogram of residuals (right)



**Analysis:**  
Residuals are tightly clustered around 0, showing the model has no systematic bias. Only minor outliers exist, which is expected for survey data.

# 4. Health Domain Interpretation

* **Clean water and healthcare access** strongly predicted better outcomes, showing the importance of infrastructure and service availability.
* **Immunization rates** are critical predictors, aligning with known public health interventions.
* **HIV indicators** remain significant, consistent with South Africa’s epidemiological profile.

# 5. Limitations and Recommendations

* **Limitations:** Small dataset (609 records), potential survey bias, and Random Forest’s black-box nature reduce interpretability.
* **Recommendations:**
  + Complement Random Forest with interpretable models (e.g., SHAP).
  + Collect larger and more recent datasets to confirm findings.
  + Use results as decision-support, not deterministic policy tools.

# 6. Conclusion

The Random Forest model demonstrated strong predictive power with stable diagnostics.  
The health domain insights align with known risk factors, reinforcing the reliability of the model for policy guidance, while highlighting areas for future data and methodological improvement.