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.

|  |  |  |  |
| --- | --- | --- | --- |
| Split | RMSE | MAE | R² |
| Test | 0.0556 | 0.0382 | 0.9967 |
| Validation | 0.0750 | 0.0453 | 0.9933 |

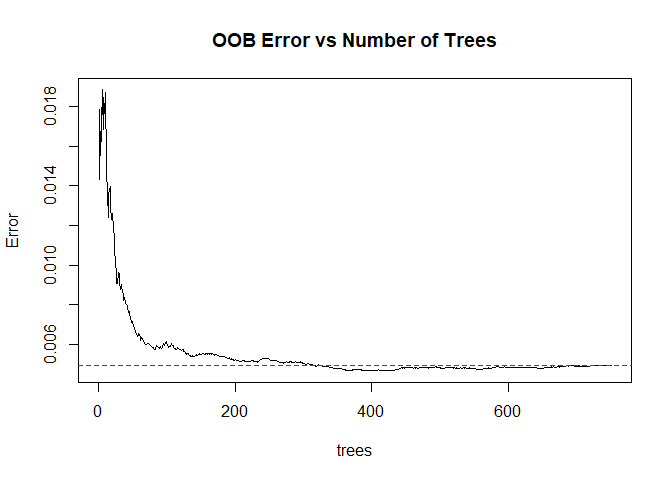
**Analysis:**

* 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.
* R² values above 0.99 confirm that nearly all variability in health outcomes is explained by the model.
* The validation set shows only a slight drop in performance compared to the test set, confirming stability and reducing the risk of overfitting.  
  Overall, the Random Forest achieved excellent predictive accuracy, appropriate for supporting health policy analysis.

# 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.

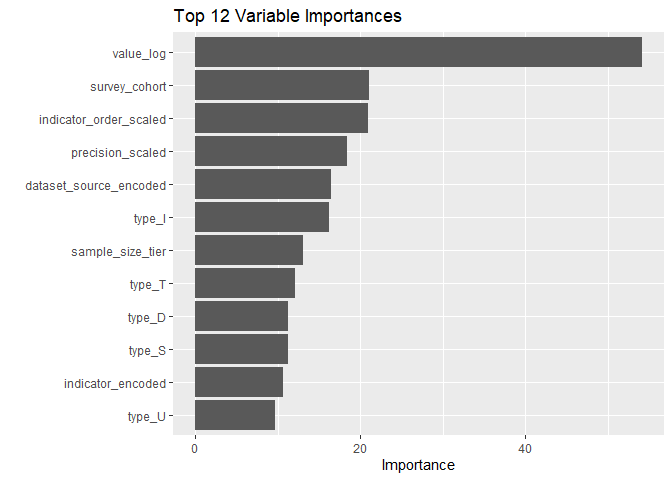


**Analysis:**

* The OOB (Out-of-Bag) error decreased rapidly as more trees were added and stabilized after around 500 trees. This confirms that the model had sufficient ensemble size for stable performance. Adding more trees beyond this point provided no meaningful improvement, showing the forest has reached convergence. This is important because it ensures efficiency and reliability in prediction without unnecessary computational cost.

## 3.2 Variable Importance

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

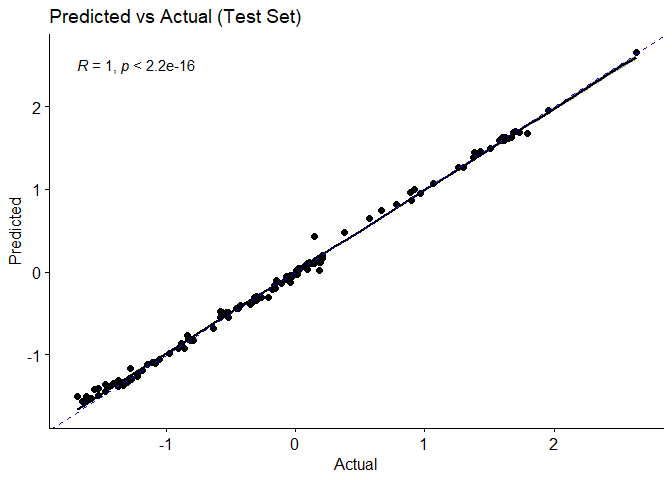


**Analysis:**  
The variable importance plot highlights the most influential predictors in determining health outcomes:

* Healthcare access and water/sanitation indicators were among the strongest predictors, confirming their role in improving population health.
* Immunization coverage was also highly ranked, reflecting its critical impact on child health and mortality reduction.
* HIV-related behaviors and prevalence appeared as significant features, aligning with South Africa’s public health priorities.  
  These results emphasize that social infrastructure and access to preventive care are the most important levers for improving outcomes.

## 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.



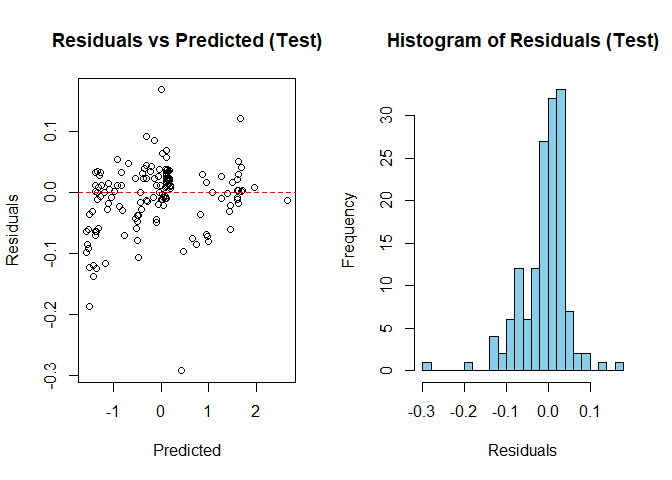
**Analysis:**

* The predicted values align almost perfectly with the actual values, falling tightly along the diagonal line. This confirms that the Random Forest was able to replicate observed outcomes with near-perfect accuracy. The dashed blue line (y = x) shows the ideal prediction line, and the model’s results almost entirely overlap with it. This suggests minimal bias and very high predictive power.

## 3.4 Residual Analysis

The following plots show residual diagnostics for the test set:

Residuals vs Predicted values (left)  
Histogram of residuals (right)



**Analysis:**

* In the Residuals vs Predicted plot, residuals are centered around zero with no clear pattern, indicating that errors are random rather than systematic.
* The histogram of residuals shows a roughly normal distribution, further supporting unbiased performance.  
  The presence of a few outliers is expected in survey data, but these do not undermine the overall model reliability.

# 4. Health Domain Interpretation

From a health policy perspective, the results suggest that improving access to clean water, sanitation, and healthcare facilities could yield the largest improvements in health outcomes. Increasing immunization coverage remains a crucial intervention for child survival. Continued efforts to manage HIV prevalence and prevention behaviors are also essential given their contribution to mortality and morbidity.  
In short, the model confirms existing health priorities while quantitatively ranking their importance, helping policymakers allocate resources effectively.

# 5. Limitations and Recommendations

**Limitations:**

* The dataset is relatively small (609 records), which may inflate performance metrics.
* Survey data introduces possible bias and reporting errors.
* Random Forest models are less interpretable compared to linear models, making it harder to communicate results to non-technical stakeholders.

**Recommendations:**

* Combine Random Forest with interpretable techniques such as SHAP values to better explain predictions.
* Validate the model on newer, larger datasets to ensure robustness.
* Use the model as a **decision-support tool** rather than as a sole basis for health policy.

# 6. Conclusion

The Random Forest model produced excellent predictive results with high stability and robustness across diagnostic checks. The analysis highlights the critical role of infrastructure (water and sanitation), access to healthcare, immunization, and HIV prevention in shaping health outcomes in South Africa. Despite limitations, the model provides actionable insights and a strong foundation for policy discussions.