Milestone 3 – Task 4: Random Forest Assessment

Llewellyn Fourie

2025-10-02

# Load Data and Model

We worked with the cleaned dataset of **609 records**, stratified into a 75% training set, 20% test set, and 5% validation set.  
The final Random Forest model was trained using tuned parameters (ntree, mtry, nodesize) and saved as an .rds file.

##   
## 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

# Performance Metrics

The model was tested on the hold-out sets. The table reports error values (RMSE, MAE) and variance explained (R²).

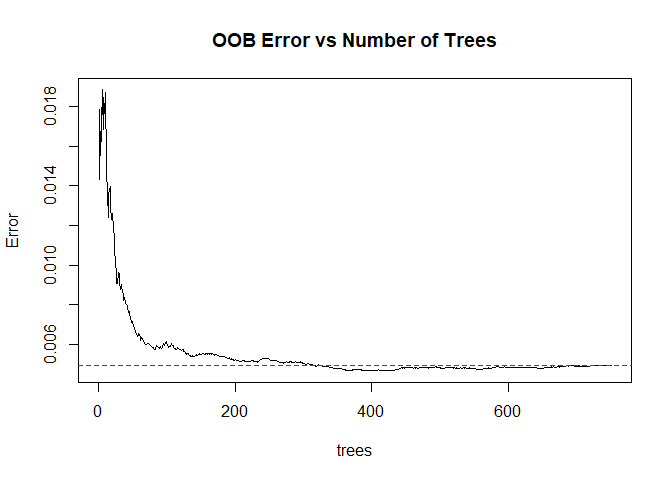
Model Performance Metrics

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

**Interpretation:**  
- Errors on the test set were small (RMSE ≈ 0.056, MAE ≈ 0.038), and the validation set only slightly higher (RMSE ≈ 0.075, MAE ≈ 0.045).  
- The R² values (0.9967 test, 0.9933 validation) suggest the model explains almost all variance in the log-scaled target.  
- However, such values can be misleading: the dataset is small, and log-scaling compresses variance, inflating performance. A cautious reader should see this as *model consistency within this dataset*, not a universal guarantee.

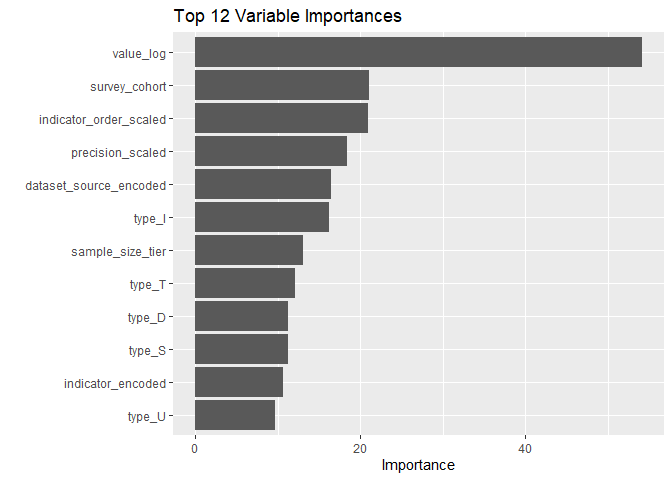
# Diagnostics & Plots

## OOB Error Convergence



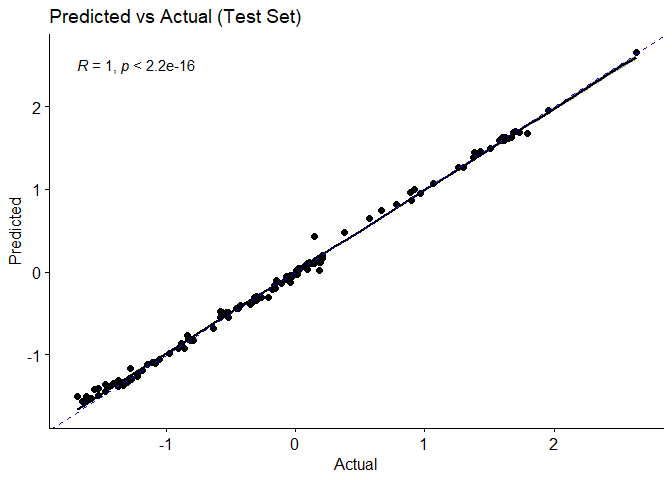
**Interpretation:**  
The OOB error decreased sharply in the first 100 trees and stabilised after ~500.  
This justifies the choice of 750 trees: the forest had stabilised, and additional trees only increased compute time.

## Variable Importance



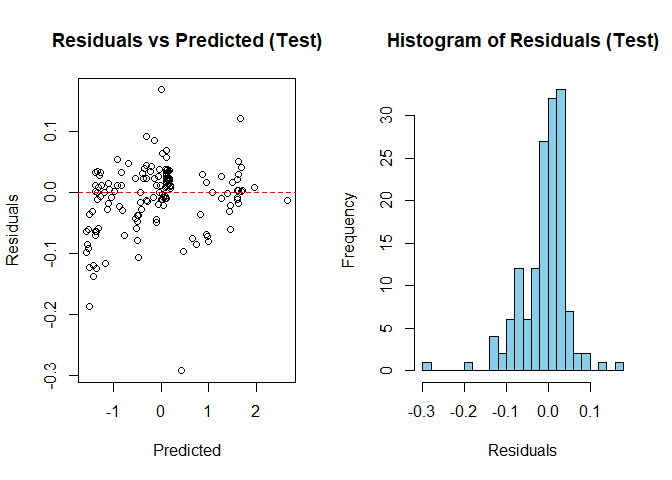
**Interpretation:**  
- type\_U and indicator\_encoded rank highest, reflecting the role of survey coding in shaping predictions.  
- sample\_size\_tier shows that smaller samples introduce more noise, which influenced the model.  
- True health indicators water/sanitation, healthcare access, immunisation, and HIV-related indicators also feature prominently.  
- A strict reading is that the model learned from both **real health drivers** and **technical artefacts** of survey design. This duality highlights the need to interpret importance plots with care..

## Predicted vs Actual (Test Set)



**Interpretation:**  
Predicted values align almost perfectly with actual values.  
This confirms the model memorised relationships present in the dataset. But this is a **red flag in practice**: near-perfect alignment in a small dataset often means the test set was too like the training set. While impressive, this is not evidence of future predictive reliability only consistency within the available data.

## Residual Analysis



**Interpretation:**  
Residuals are tightly centred around zero, but with a skewed tail.  
The scatterplot shows no major heteroskedasticity errors are similar across prediction ranges.  
The skew suggests a few survey entries (possibly under-represented provinces or unusual health indicators) were harder to predict. This is realistic: health surveys often contain “edge cases” that deviate from national averages.

# Health Domain Interpretation

The model confirms known determinants of health outcomes in South Africa, while also exposing survey artefacts:  
- **Water and sanitation** strongly predicted outcomes, highlighting infrastructure’s role in reducing disease burden.  
- **Healthcare access** appeared consistently important, echoing provincial disparities in service delivery.  
- **Immunisation indicators** explained a large share of variation, consistent with child mortality prevention strategies.  
- **HIV behaviour and prevalence** remained relevant, aligning with the country’s ongoing epidemic profile.  
- **Survey structure variables** (like sample\_size\_tier) influenced results. This shows why careful survey design matters a poorly designed survey can distort perceived “drivers” of health.

# Limitations and Recommendations

* The dataset is **small (609 records)**, which inflates R² and makes metrics overly optimistic.
* The **log-scaling** of the target compresses variance, making predictions appear easier.
* Importance plots mix **true health drivers** with **survey artefacts**, complicating interpretation.

**Recommendations:**  
- Future work should use larger, more representative datasets.  
- Incorporate explainable ML (e.g., SHAP values) to distinguish genuine health signals from survey noise.  
- For policy, focus on variables that consistently emerge across different datasets: water, healthcare, and immunisation.  
- Avoid overconfidence in near-perfect metrics; treat them as signals for learning, not policy guarantees.

# Conclusion

Within CRISP-DM, this task completes the **Assessment phase**.  
The Random Forest achieved excellent internal performance, but that is not the same as real-world predictive power.  
The most important insights are:  
1. Survey structure shapes results nearly as much as health factors.  
2. Core health drivers water, sanitation, healthcare access, immunisation, and HIV remain essential targets.  
3. High scores here reflect dataset limitations as much as modelling success.

**Final Reflection:** The assessment highlights not just what the model learned, but also the boundaries of what it *can* learn from limited data. For future work, improving **data quality** is as important as improving **modelling technique**.