HDPSA Milestone 4

Evaluation Report

Bin 381 Group A

10/10/2025

Members:

Llewellyn Jakobus Fourie – 601314

Juan Oosthuizen – 600161

Erin David Cullen – 600531

Qhamaninande Ndlume – 601436

Table of Contents

Executive Summary

This evaluation report represents the culmination of Phases 1-4 of the CRISP-DM methodology applied to predicting health outcomes in South Africa using a Random Forest regression model. The report addresses three critical evaluation tasks mandated by Milestone 4:

1. **Evaluate Results** - Assessment of model performance against business objectives
2. **Review the Process** - Comprehensive audit of CRISP-DM execution across all phases
3. **Determine Next Steps** - Evidence-based recommendation for project progression

Key Findings

Model Performance

The Random Forest model achieves exceptional performance metrics that exceed all predefined business success criteria:

- R-squared = 0.997 (99.7% variance explained) - exceeds 0.75 target by 32.9%
- RMSE = 0.0554 - well below 0.10 threshold
- MAE = 0.0381 - below 0.05 threshold
- Feature importance aligns with policy priorities: Water Access, Sanitation, Literacy, Healthcare Access

Process Quality Assessment

The process review identified **3 critical issues**, **5 major issues**, and **11 minor issues** across the four CRISP-DM phases:

- **Critical concerns** include potential data leakage from scaling procedures, test set contamination during hyperparameter tuning, and the exceptionally high R-squared requiring validation
- **Strengths** include systematic documentation, appropriate model selection, and policy-aligned interpretability

Final Recommendation

Proceed with a focused two-week validation iteration to address critical methodological concerns, followed by controlled pilot deployment contingent on meeting acceptance criteria. This balanced approach prioritizes methodological rigor while enabling timely policy impact.

Task 1: Evaluate Results

Assessment of Random Forest Model Results

Re-establishing Business Goals (from Milestone 1)

The model's performance is evaluated against the four business goals established in Milestone 1:

Business Goal	Success Criterion / Threshold	Evidence from Random Forest Model	Quantitative Result	Evaluation
BG1: Predict national and provincial health outcomes to support policy intervention	Model accuracy >= 70% or R-squared > 0.75	R-squared = 0.997 (99.7% variance explained)	+32.9% above target	Achieved
BG2: Maintain low prediction error for continuous health indicators (Value variable)	RMSE <= 0.10 and MAE <= 0.05 (on log-scaled health index)	RMSE = 0.0554, MAE = 0.0381	Both below threshold	Achieved
BG3: Provide interpretable outputs for stakeholders (Gov / NGOs / DoH)	Top predictors must align with key health determinants	Feature importance: (1) Water Access, (2) Sanitation, (3) Literacy, (4) Healthcare Access	4 / 4 policy-relevant drivers identified	Achieved
BG4: Ensure robustness and ethical reliability on limited data (609 records)	Minimal overfitting (OOB MSE approximately equal to Test MSE) and stable residuals	OOB MSE = 0.0049 approximately equals Test MSE (0.0031); Delta = 0.0018 (< 0.01)	Stable and generalises well	Achieved

All four business goals have been met or exceeded, demonstrating strong alignment between technical performance and stakeholder requirements.

Interpretation of Metrics vs Success Thresholds

Root Mean Squared Error (RMSE = 0.0554)

On a log-scaled 0-1 health index, this indicates a 5.5% average deviation from actual values, comfortably within the acceptable 10% margin for reliable policy modelling. This level of precision supports confident resource allocation and intervention prioritization decisions.

Mean Absolute Error (MAE = 0.0381)

Represents a 3.8% average absolute deviation, surpassing the desired threshold for national-level health indicator predictions. The MAE metric is particularly interpretable for stakeholders, as it represents the typical prediction error in the original scale units.

Coefficient of Determination (R-squared = 0.997)

The model explains 99.7% of the total variance, exceeding the defined target (R-squared > 0.75) by a substantial margin. Out-of-bag error (0.0049) and test RMSE-squared (0.0031) are nearly identical, confirming minimal overfitting and high generalisation capacity. This exceptional performance warrants careful validation to ensure it reflects genuine predictive capability rather than methodological artifacts.

Out-of-Bag Validation

The Random Forest's internal out-of-bag (OOB) validation mechanism provides an unbiased performance estimate without requiring a separate validation set. The OOB MSE of 0.0049 aligns closely with test set performance, providing evidence of model stability and generalization capability.

Critical Reflection and Policy Relevance

Feature Importance and Policy Alignment

Feature importance analysis identifies water access and sanitation as the primary drivers of health outcomes, ranking first and second respectively among all predictors. This aligns directly with national priorities related to clean water, sanitation and sustainable development (SDG 6). The prominence of literacy and healthcare access as secondary drivers further validates the model's capture of established health determinants documented in public health literature.

Addressing the High R-squared

Although the exceptionally high R-squared might suggest potential overfitting, the alignment between OOB error and test performance provides supporting evidence for generalization. However, as identified in Task 2, this exceptional performance warrants methodological validation to rule out data leakage or test set contamination that could artificially inflate metrics.

Actionable Policy Insights

The model's feature importance rankings provide clear priorities for policy intervention, with water access and sanitation emerging as the primary drivers of health outcomes. These findings support investment in infrastructure improvements aligned with SDG 6 (Clean Water and Sanitation). Future work should include partial dependence analysis to quantify specific effect sizes and enable precise cost-benefit calculations for intervention planning.

Ethical Considerations

Ethically, the dataset contains no personal identifiers, ensuring compliance with privacy standards. However, underrepresentation from rural regions remains a limitation that may affect predictive balance across geographic subgroups. Future iterations should assess model performance stratified by province and urban-rural status to detect potential biases.

Model Approval Recommendation

The Random Forest regression model **meets and exceeds all predefined business success criteria**. Its exceptional accuracy (R-squared = 0.997), low error rates (RMSE = 0.055; MAE = 0.038), and policy-aligned feature interpretability demonstrate strong predictive validity and operational value.

Conditional Approval: The model is recommended for advancement to validation iteration, with deployment contingent on addressing the critical methodological concerns identified in Task 2. Specifically:

1. Verification of no data leakage in preprocessing pipeline
2. Confirmation that hyperparameter tuning did not contaminate test set
3. Validation of performance against simple baseline models

Given its robustness, reliability and transparency, the model shows strong potential for adoption in national and provincial health policy planning. Continuous monitoring and future retraining with expanded datasets are advised to maintain fairness and adaptability over time.

Task 2: Review the Process

Review of the CRISP-DM Process (Phases 1-4)

Executive Summary

This review evaluates the execution of Phases 1 through 4 of the CRISP-DM methodology applied to the South African health outcomes prediction project. Overall, the project demonstrated strong adherence to CRISP-DM principles with well-documented processes, appropriate methodological choices, and excellent technical outcomes. However, several quality assurance concerns and process gaps were identified that warrant attention for future iterations.

The review follows a structured format for each phase: properly executed steps, issues identified, quality assurance concerns, and corrective actions recommended.

Phase 1: Business Understanding

Properly Executed Steps

Business Objectives Definition

- Four distinct business goals were clearly articulated with measurable success criteria
- BG1: Predict health outcomes with model accuracy >= 70% or R-squared > 0.75
- BG2: Maintain low prediction error (RMSE <= 0.10, MAE <= 0.05)
- BG3: Provide interpretable outputs for government/NGO stakeholders
- BG4: Ensure robustness on limited data (609 records)

Stakeholder Identification

- Primary stakeholders clearly identified: Government agencies, NGOs, Department of Health
- Their requirements for interpretability and policy-actionable insights documented

Success Criteria Establishment

- Quantitative thresholds defined for all business goals
- Metrics aligned with health policy decision-making needs
- Ethical considerations (privacy, fairness) acknowledged

Issues Identified

Gap 1: Incomplete Risk Assessment

- While constraints were mentioned (limited data: 609 records), no formal risk register was documented
- Missing: specific mitigation strategies for small sample size risks
- Missing: data availability risks or contingency plans if datasets proved inadequate

Gap 2: Limited Stakeholder Validation

- No evidence that success criteria (70% accuracy, R-squared > 0.75) were validated with actual stakeholders
- Thresholds appear to be internally defined rather than derived from stakeholder requirements
- Recommendation: Future projects should include stakeholder interviews to confirm acceptance criteria

Gap 3: Scope Boundary Ambiguity

- Project scope does not explicitly state what is OUT of scope
- Unclear whether provincial-level predictions vs. national-level predictions were prioritized
- Missing: timeline constraints and resource allocation details

Quality Assurance Concerns

Concern 1: Business Success Metrics May Be Too Lenient

- 70% accuracy threshold is relatively low for health policy applications where errors affect resource allocation
- No justification provided for why this threshold is appropriate
- Actual model performance (R-squared = 0.997) far exceeds this, suggesting criteria could have been more ambitious

Concern 2: No Business ROI Analysis

- Missing cost-benefit analysis of model deployment
- No estimation of potential policy impact or cost savings from improved predictions
- Future projects should quantify expected business value

Corrective Actions Recommended

1. **Develop formal risk register** documenting data risks, technical risks, and business risks with mitigation plans
2. **Validate success criteria with stakeholders** through interviews or workshops before modeling phase
3. **Define explicit scope boundaries** including geographic coverage, time horizons, and excluded indicators
4. **Add business value quantification** to justify project investment

Phase 2: Data Understanding

Properly Executed Steps

Data Collection

- Successfully gathered 13 health and demographic datasets from South Africa
- Datasets cover diverse health indicators: access to healthcare, child mortality, immunization, water access, sanitation, literacy, HIV behavior
- Total of 609 records after integration

Initial Data Exploration

- Basic descriptive statistics computed for key variables
- Data structure documented (rows, columns, data types)
- Missing value percentages calculated for all datasets

Data Quality Assessment

- Systematic check for missing values across all datasets
- Duplicate detection performed
- Numeric correlation analysis conducted to identify multicollinearity

Issues Identified

Gap 4: Insufficient Data Profiling

- Limited documentation of value distributions (skewness, kurtosis, outliers)
- No visual exploration documented in Milestone 1 (histograms, box plots, scatter plots)
- Correlation analysis mentioned but results not fully documented in final report

Gap 5: No Temporal Analysis

- Datasets likely contain temporal dimensions (survey years, time periods)
- No analysis of time trends or temporal consistency documented
- Missing: assessment of whether data from different years can be safely combined

Gap 6: Limited External Validation

- No comparison of dataset statistics against known population benchmarks
- Example: Are reported child mortality rates consistent with WHO/Stats SA published figures?
- Missing sanity checks that would catch data entry errors or miscoded values

Quality Assurance Concerns

Concern 3: Data Representativeness Not Verified

- 609 records may not represent all provinces equally
- No documentation of geographic coverage or population representativeness
- Risk: Model may perform poorly on underrepresented regions

Concern 4: Variable Relationships Underexplored

- While correlation was mentioned, no evidence of deeper relationship analysis
- Missing: scatter plots, pair plots, or domain-specific hypothesis testing
- Example: Was the expected negative correlation between water access and child mortality confirmed?

Corrective Actions Recommended

1. **Add comprehensive visual EDA** including distribution plots, correlation heatmaps, and outlier detection visualizations
2. **Conduct temporal analysis** to verify data can be safely aggregated across time periods
3. **Perform external validation** by comparing key statistics against published health reports
4. **Document geographic representativeness** by analyzing provincial coverage

Phase 3: Data Preparation

Properly Executed Steps

Data Selection

- 7 out of 13 datasets selected based on relevance, accessibility, and manageability
- Selected datasets: access-to-health-care, immunization, hiv-behavior, water, dhs-quickstats, toilet-facilities, child-mortality-rates
- Final dataset: 609 records, 11 features for modeling

Data Cleaning

- Duplicate removal implemented systematically
- Missing value imputation strategy defined:
    - Numeric variables: median imputation (robust to outliers)
    - Categorical variables: "Unknown" category
    - Boolean variables: modal imputation
- Guidelines established: Drop columns with >40% missing values

Feature Engineering

- Categorical encoding implemented (indicator_encoded, survey_cohort, dataset_source_encoded)
- Rare category grouping applied (threshold: 5% of records)
- Dummy variable creation for categorical features
- Log transformation applied to skewed target variable (value_log_scaled)

Data Transformation

- Numeric variables scaled using standardization (z-scores)
- Created derived features: high_precision, char_order_quintile, data_quality_score
- Sample size tiering (Small/Medium/Large) for stratification

Train-Test Split

- 75% training (457 records)
- 20% testing (122 records)
- 5% validation (30 records)
- Random seed set (42) for reproducibility

Issues Identified

Gap 7: Inconsistent Train-Test Split Documentation

- Milestone 2 documentation mentions 70/30 split
- Milestone 3 implementation uses 75/20/5 split
- No explanation provided for this change
- Risk: Confusion during replication or auditing

Gap 8: Missing Value Imputation Not Validated

- Median/mode imputation applied but no analysis of impact on distributions
- No comparison of pre/post imputation statistics
- Risk: Imputation may introduce bias that affects model performance
- Missing: sensitivity analysis to assess whether imputation strategy affects model outcomes

Gap 9: Feature Selection Process Undocumented

- Final model uses 11 features but rationale for feature inclusion/exclusion not documented
- No feature importance analysis during preparation phase
- No documentation of which features from the 7 datasets were retained vs. dropped
- Recommendation: Document systematic feature selection using correlation, VIF, or domain knowledge

Gap 10: No Data Leakage Prevention (CRITICAL)

- Scaling was performed on combined data before splitting (based on Milestone 2/3 code review)
- Proper approach: Fit scaler on training data only, then transform test/validation sets
- Risk: Test set statistics leak into training process, inflating performance metrics
- Critical Issue: This may partially explain the exceptionally high R-squared (0.997)

Quality Assurance Concerns

Concern 5: Potential Data Leakage (CRITICAL)

- If scaling was performed before train-test split, test set mean/variance influenced training data normalization
- This violates the independence assumption and leads to optimistically biased metrics
- Recommendation: Verify scaling order; re-run if leakage detected

Concern 6: Small Validation Set (30 records)

- 5% validation set (30 records) may be too small for reliable final model assessment
- High variance in validation metrics expected with such limited data
- Alternative: Use k-fold cross-validation exclusively or increase validation set to 10-15%

Concern 7: Feature Engineering Complexity Not Justified

- Extensive feature engineering (dummy variables, tiering, quintiles) but no ablation study
- Unclear whether added complexity improves model or introduces noise
- Recommendation: Compare simple vs. complex feature sets to validate engineering choices

Corrective Actions Recommended

1. **CRITICAL: Re-verify scaling procedure** to ensure no data leakage occurred
    - If leakage detected, re-fit scaler on training data only and re-evaluate model
2. **Document train-test split rationale** and ensure consistency across all documentation
3. **Validate imputation strategy** by comparing model performance with/without imputation
4. **Increase validation set size** to 10-15% or rely solely on cross-validation
5. **Add feature selection documentation** explaining which features were retained and why
6. **Conduct ablation study** to validate feature engineering contributions

Phase 4: Modeling

Properly Executed Steps

Model Selection and Justification

- Random Forest Regression selected with strong justification:
  - Handles mixed data types (categorical + numerical)
  - Robust to outliers and missing data
  - Provides variable importance for interpretability
  - No distribution assumptions required
  - Built-in OOB validation
- Alternative models considered (Linear Regression, XGBoost, Neural Networks) with documented rationale for rejection

Test Design

- Clear evaluation metrics defined: RMSE, MAE, R-squared
- Random Forest-specific metrics included: OOB error, variable importance
- 5-fold cross-validation implemented for robust performance estimation
- Systematic hyperparameter tuning approach designed

Model Building

- Comprehensive hyperparameter tuning performed:
  - ntree tuned: 500-2000 (optimal: 750)
  - mtry tuned: 5-9 (optimal: 9)
  - nodesize tuned: 1-10 (optimal: 2)
- Sequential tuning approach (ntree -> mtry -> nodesize) clearly documented
- Each tuning step evaluated on test RMSE and OOB error
- Final model parameters well-justified with quantitative evidence

Model Assessment

- Excellent performance metrics achieved:
  - Test RMSE: 0.0554
  - Test MAE: 0.0381
  - R-squared: 0.997 (99.7% variance explained)
  - OOB MSE: 0.0049
- Variable importance analysis identifies policy-relevant features:
  1. Water Access
  2. Sanitation
  3. Literacy
  4. Healthcare Access
- Cross-validation stability confirmed (low variance across folds)

Issues Identified

Gap 11: Hyperparameter Tuning Used Test Set (CRITICAL)

- Tuning process evaluated models on the test set at each step
- Test set should be held out until final evaluation only
- Proper approach: Use cross-validation on training set for tuning, test set for final assessment
- Risk: Tuning on test set causes overfitting to test data, inflating performance estimates

Gap 12: No Baseline Model Comparison

- Random Forest compared against "baseline RF with default parameters" but not against simpler models
- Missing: Performance comparison with linear regression or mean/median baseline
- Difficult to assess true value-add of Random Forest without simple baseline

Gap 13: Limited Residual Analysis

- Task 1 mentions "residual diagnostics" but no plots or detailed analysis provided
- Missing: residual plots, Q-Q plots, heteroscedasticity tests
- Missing: analysis of where model fails (which records have highest errors?)

Gap 14: No Model Interpretability Deep-Dive

- Variable importance reported but no partial dependence plots or SHAP values
- Missing: quantified effect sizes showing how features affect predictions
- Missing: concrete interpretations of marginal effects (e.g., impact of 1% increase in water access)
- Stakeholders need actionable insights beyond feature rankings

Gap 15: Limited Cross-Validation Documentation

- Model relies primarily on OOB validation and single train-test split
- No k-fold cross-validation implemented to assess performance stability across multiple partitions
- Missing: Cross-validation analysis to validate robustness beyond OOB estimates

Quality Assurance Concerns

Concern 8: Suspiciously High R-squared (0.997) (CRITICAL)

- 99.7% variance explained is exceptionally rare in real-world regression tasks
- Possible explanations:
    1. Data leakage (scaling before split, tuning on test set)
    2. Target variable included in features (e.g., value_log used as both target and feature)
    3. Multicollinearity causing overfitting
    4. Genuine excellent fit (least likely for health survey data)
- Recommendation: Audit feature matrix to ensure target variable not inadvertently included

Concern 9: OOB vs. Test RMSE Discrepancy

- OOB MSE: 0.0049 (equivalent RMSE: 0.070)
- Test RMSE: 0.0554
- While close, OOB error is higher than test error, which is unusual
- Typically test error >= OOB error due to generalization gap
- Suggests potential test set contamination or anomaly

Concern 10: Small Test Set (122 records)

- Test set contains only 122 records
- High variance expected in test metrics
- Single train-test split may not be representative
- Recommendation: Report confidence intervals for test metrics or use nested cross-validation

Concern 11: Variable Importance Not Validated

- Feature importance based on single model run
- No stability analysis (e.g., do top features remain consistent across CV folds?)
- Risk: Rankings may be artifacts of specific train-test split

Corrective Actions Recommended

1. **CRITICAL: Audit for data leakage**
    - Verify target variable (value_log_scaled) is not present in feature matrix
    - Check scaling was performed after train-test split
    - Re-run model with proper hold-out procedures
2. **Re-implement hyperparameter tuning using cross-validation only**
    - Reserve test set for final evaluation
    - Use nested CV for unbiased hyperparameter selection
3. **Add baseline model comparisons**
    - Train linear regression, mean predictor, median predictor
    - Report performance gaps to quantify Random Forest value-add
4. **Conduct residual analysis**
    - Plot residuals vs. fitted values
    - Identify high-error cases for investigation
    - Check for heteroscedasticity or systematic biases
5. **Add model interpretability analysis**
    - Partial dependence plots for top 4 features
    - SHAP values or permutation importance for validation
    - Derive concrete policy insights (e.g., effect sizes)
6. **Validate variable importance stability**
    - Compute importance across all CV folds
    - Report mean importance and variance

Summary of Quality Assurance Issues

Critical Issues (Must Address)

1. **Potential data leakage from scaling before split** (Phase 3)
2. **Hyperparameter tuning performed on test set** (Phase 4)
3. **Suspiciously high R-squared (0.997) requires investigation** (Phase 4)

Major Issues (Should Address)

4. **Train-test split inconsistency** (70/30 vs. 75/20/5) (Phase 3)
5. **Small validation set (30 records)** (Phase 3)
6. **Missing value imputation not validated** (Phase 3)
7. **No baseline model comparison** (Phase 4)
8. **Limited residual diagnostics** (Phase 4)

Minor Issues (Nice to Have)

9. **Incomplete risk assessment** (Phase 1)
10. **Limited stakeholder validation of success criteria** (Phase 1)
11. **Insufficient data profiling** (Phase 2)
12. **No temporal analysis of datasets** (Phase 2)
13. **Feature selection process undocumented** (Phase 3)
14. **No model interpretability deep-dive** (Phase 4)
15. **Variable importance stability not validated** (Phase 4)

Lessons Learned

What Worked Well

1. **CRISP-DM methodology provided clear structure** - Each phase built logically on the previous one
2. **Comprehensive documentation** - Most decisions were recorded with rationale
3. **Systematic hyperparameter tuning** - Sequential optimization approach was methodical and well-documented
4. **Reproducibility enabled** - Random seeds, file paths, and code structure support replication
5. **Domain-aligned feature importance** - Results align with known health determinants (water, sanitation)

What Could Be Improved

1. **Data leakage prevention protocols** - Need stricter separation between train/test/validation throughout pipeline
2. **Baseline comparisons** - Always compare complex models against simple baselines to demonstrate value
3. **Test set discipline** - Reserve test set exclusively for final evaluation; use CV for all tuning
4. **Stakeholder engagement** - Involve stakeholders earlier to validate success criteria and interpretability needs
5. **Audit trails for data transformations** - Document exactly which transformations were applied when and why
6. **Sanity checks at every phase** - Implement automated checks (e.g., target variable not in features, scaling order correct)

Recommendations for Future Projects

Process Improvements

1. **Implement data leakage checklist**
    - Verify: Scaling fit on training data only
    - Verify: Test set never used for hyperparameter tuning
    - Verify: Target variable not in feature matrix
    - Verify: Temporal ordering preserved (if time-series data)
2. **Standardize train-test-validation protocol**
    - Document split ratios in project charter (Phase 1)
    - Implement splits at start of Phase 3 and never change
    - Use stratified sampling when appropriate
    - Consider nested cross-validation for small datasets
3. **Add baseline model requirement**
    - Always train at least one simple baseline (mean, median, linear regression)
    - Report performance deltas to justify complex model choices
    - Include in Milestone 3 deliverables
4. **Enhance interpretability analysis**
    - Require partial dependence plots for top features
    - Include SHAP values or permutation importance
    - Translate feature importance into actionable policy insights
    - Make this a required section in Milestone 3
5. **Strengthen Phase 1 stakeholder validation**
    - Conduct stakeholder interviews before defining success criteria
    - Document stakeholder requirements explicitly
    - Validate model outputs with stakeholders before deployment decision
6. **Add automated quality checks**
    - Implement unit tests for data pipelines
    - Add assertions to catch data leakage (e.g., assert train/test indices disjoint)
    - Check for target variable in feature columns
    - Validate distribution shifts between train/test sets

Technical Recommendations

1. **Use pipeline objects** (e.g., scikit-learn Pipeline, R caret) to enforce correct transformation order
2. **Report confidence intervals** for all test metrics using bootstrap or CV
3. **Implement nested cross-validation** for hyperparameter tuning on small datasets
4. **Add residual analysis** as standard deliverable in modeling phase
5. **Conduct sensitivity analysis** to assess robustness to imputation and scaling choices

Documentation Recommendations

1. **Create audit trail document** tracking all data transformations with timestamps
2. **Maintain decision log** recording key choices and rationale
3. **Add reproducibility checklist** ensuring code can be re-run from scratch
4. **Include data dictionaries** defining all variables and transformations
5. **Document lessons learned** at end of each phase, not just final evaluation

---

Process Review Conclusion

The project demonstrated strong technical execution with excellent model performance and adherence to CRISP-DM structure. However, several critical quality assurance concerns were identified, particularly around potential data leakage and test set contamination during hyperparameter tuning.

The exceptionally high R-squared (0.997) warrants investigation to rule out methodological errors. While the model may genuinely perform well, such extreme performance on limited health survey data is unusual and should be validated through corrective actions.

Despite these concerns, the project provides a solid foundation for deployment pending resolution of critical issues. The systematic approach, comprehensive documentation, and policy-relevant insights demonstrate the value of CRISP-DM methodology for health analytics projects.

Overall Assessment

Phase	Execution Quality	Critical Issues	Major Issues	Minor Issues
Phase 1: Business Understanding	Good	0	0	3
Phase 2: Data Understanding	Moderate	0	0	4
Phase 3: Data Preparation	Moderate	1	3	1
Phase 4: Modeling	Good	2	2	3
TOTAL	**Good**	**3**	**5**	**11**

Process Review Recommendation: Address 3 critical issues before deployment. Model shows strong potential but requires methodological validation to ensure results are not artifacts of data leakage or test set contamination.

Task 3: Determine the Next Steps

Context Recap

This project applies the CRISP-DM methodology to the Health and Demographic Profile of South Africa (HDPSA), developing a Random Forest regression model to predict a log-scaled health outcome (value_log_scaled). The evidence base supporting decisions at this Evaluation phase comprises three elements, each serving a distinct function and each essential to a defensible decision about deployment or iteration.

Evidence Base

- **Scaled data for modelling.** The final engineered dataset in the Scaled Data folder consolidates cleaned demographic and health indicators into approximately 31 numeric and categorical features. These features capture core policy-relevant domains - water access, sanitation, literacy, and healthcare access - prepared for learning through centering/scaling and encoding where appropriate. This table defines the feature space and thus determines both the stability and interpretability of the model.
- **Reproducible data splits.** The Split Data folder contains the training, validation, and test partitions. These partitions ensure consistent, reproducible estimation of performance, provided that the test set remains untouched until final evaluation and that all transformation steps are fitted on the training subset only. This discipline protects against optimistic bias and guards model credibility.
- **Modelling outputs and artefacts.** The Milestone 3 outputs include hyperparameter tuning logs (e.g., for mtry and node size) and a final model artefact. Together they document how the model was selected, with final parameters near mtry approximately equals 9, nodesize approximately equals 2, and ntree = 750. These artefacts also make transparent any assumptions that must be validated in this milestone, including the risk of data leakage if preprocessing was fit before the train-validation-test separation or if the test set guided tuning.

Performance Summary

Milestone 3 reported strong predictive performance on the log scale (RMSE = 0.0554; MAE = 0.0381; R-squared = 0.997) on a sample of roughly 600 records. Feature rankings foreground water and sanitation access, literacy, and healthcare access - determinants aligned with health-policy priorities. The present decision focuses on whether these results are sufficiently robust, explainable, and ethically appropriate to justify deployment now or after a short iteration.

Options Analysis

Option 1: Deploy (Production or Limited Pilot)

Rationale

The model currently meets stringent thresholds on the selected scale and offers clear policy relevance by surfacing interpretable drivers. A limited pilot would enable controlled, real-world validation while informing resource allocation.

Benefits

- Accelerates value capture
- Establishes monitoring in an operational setting
- Leverages existing artefacts with minimal rework

Constraints

Without a brief validation cycle, deployment risks propagating optimistic metrics if leakage or inadvertent test-set tuning occurred. Given the modest dataset size, external validity must be demonstrated under unbiased estimation procedures.

Safeguards

- Pre-deployment audit of split integrity and transformation order
- Ethical review and subgroup checks
- Monitoring with rollback controls

Option 2: Iterate (Short Validation Cycle) - RECOMMENDED

Rationale

A two-week, time-boxed iteration will close methodological risks and produce unbiased estimates via cross-validation, confirm separation of train/validation/test, quantify the model's value-add versus simple baselines, and deepen interpretability.

Benefits

- Strengthens credibility
- Reduces risk of optimistic bias
- Provides robust evidence for stakeholders
- Improves transparency via diagnostics and stability analyses

Constraints

- Short delay to deployment
- Limited additional analytical effort

Safeguards

- Strict test-set holdout
- Pipeline-based preprocessing fitted on training data only
- Full audit trail of transformations and decisions

Option 3: Abandon/Restart

Rationale

Consider only if leakage is confirmed and cannot be mitigated with current data, or if project objectives materially change.

Drawbacks

Current results are promising and policy-aligned; abandoning would discard a strong foundation for impact.

Conditions

- Irreparable data contamination
- Significant scope change
- Sustained misalignment with stakeholder needs

Final Recommendation

Proceed with a short, focused iteration to validate methodology, then advance to a controlled pilot if acceptance criteria are met on an untouched hold-out set. This pathway balances policy urgency with methodological rigor and ethical assurance, ensuring that deployment decisions rest on defensible, transparent evidence.

Acceptance criteria after the validation cycle:

- RMSE <= 0.06, MAE <= 0.04, R-squared >= 0.95 on the log-scaled target
- Stable top predictors across folds
- No evidence of leakage
- Simple baselines underperform the Random Forest by a meaningful margin

---

Quantitative Evidence

The current model performance meets all proposed acceptance thresholds:

Metric	Target Threshold	Milestone 3 Final	Meets Goal?
RMSE	<= 0.06	0.0554	Yes
MAE	<= 0.04	0.0381	Yes
R-squared	>= 0.95	0.997	Yes

These metrics justify advancing to a short iteration prior to pilot deployment, with the caveat that methodological validation must confirm these results are not artifacts of data leakage or test contamination.

Traceability to Business Goals

Predictive accuracy for prioritization

R-squared of 0.997 exceeds the 0.70 benchmark, supporting precise identification of high-risk provinces for targeted intervention.

Operational reliability via low error

RMSE of 0.0554 and MAE of 0.0381 meet thresholds consistent with policy planning and scenario analysis.

Stakeholder interpretability

Feature rankings emphasize water, sanitation, literacy, and healthcare access - determinants that align with policy levers and enable transparent communication.

Robustness under data constraints

The forthcoming iteration will confirm generalization under strict separation of data partitions and proper transformation order.

Action Plan

Phase A: Audit and Rebuild (Week 1)

Objectives

- Validate partition integrity
- Enforce preprocessing discipline
- Re-tune via cross-validation

Activities

- Confirm disjoint indices and immutability of training, validation, and test splits; document checks and outcomes
- Ensure centering/scaling and encoding are fitted on training data only and then applied to validation/test without refitting
- Replace any test-guided tuning with k-fold cross-validation; reserve the test set strictly for the final estimate

Deliverables

- Audit report
- Updated tuning logs
- Revised model artefact
- Documented transformation parameters

Assigned to: Llewellyn Fourie (QA Audit - Leakage checks, split integrity, and data validation)

Phase B: Re-validate and Compare (Week 2)

Objectives

- Finalize Random Forest under validated procedures
- Establish baselines
- Produce diagnostics and interpretability analysis

Activities

- Produce test-set estimates on the untouched hold-out
- Train simple comparators (mean/median predictor and linear regression) and report performance deltas to demonstrate value-add
- Provide residual plots, error summaries, and permutation- or LIME-based explanations to evidence accuracy and explainability

Deliverables

- Metrics table with acceptance decision
- Diagnostics pack
- Interpretability summary
- Updated risk register

Assigned to: Juan Oosthuizen (Re-Validation - Cross-validation, baselines, diagnostics, and acceptance decision)

Phase C: Pilot Deployment (Week 3, conditional)

Objectives

- Package model for deployment
- Implement monitoring infrastructure
- Obtain stakeholder sign-off

Activities

- Version the model artefact with metadata, schema, and usage constraints; define rollback plan and access controls
- Implement metric tracking (RMSE/MAE/R-squared), data drift alerts, and periodic backtesting; schedule governance reviews

Deliverables

- Deployment checklist
- Monitoring and escalation plan
- Stakeholder sign-off

Assigned to: Erin Cullen and Qhamaninande Ndlume (Pilot - Packaging, deployment controls, monitoring, and rollback readiness)

Implementation Notes

Pipelines and guards

Use a pipeline-oriented workflow to guarantee transformation order and separation of concerns; add assertions for disjoint indices and exclusion of target-derived signals from features.

Data contracts

Fix the schema (types, ranges, categorical levels) and define handling rules for missing or novel values to ensure stable inference.

Interpretability

Report permutation importance and partial dependence/ICE visualizations for top predictors to provide stable, policy-aligned explanations.

Calibration

Where appropriate, apply post-hoc calibration to improve alignment between predicted and observed scales, documented with before/after metrics.

Risks and Mitigations

Potential data leakage

Mitigate via pipeline-enforced preprocessing on training only; audit and log all transformation fits; re-estimate metrics under validated procedures.

Test-set contamination

Reserve test data exclusively for the final estimate; perform all tuning and selection within cross-validation.

Small sample size

Use cross-validation with uncertainty estimates; monitor post-deployment performance and drift; retrain on a cadence aligned with data refreshes.

Distribution shift

Compare training and pilot-period covariate distributions; define triggers for retraining or model fallback.

Ethics and Bias

Conduct subgroup performance and explanatory stability checks across province, gender, and urban-rural strata. If systematic under- or over-prediction is detected for specific subgroups, mitigation will include recalibration, feature review, or constrained modelling choices. Explanations should be communicated in plain language to support equitable, accountable policy use.

Reproducibility and Version Control

Version all scripts, data snapshots, and model artefacts with tagged commits; record environments via session information to ensure reproducibility. Each result must be traceable to an exact configuration, dataset snapshot, and decision log, consistent with CRISP-DM documentation standards.

Confidence Intervals

Report 95% confidence intervals for test-set metrics using bootstrap resampling of residuals or cross-validated standard errors. These intervals provide decision-makers with uncertainty bounds appropriate for small datasets and guard against over-interpretation of point estimates.

CRISP-DM Alignment

These steps operationalize the Evaluation phase by validating results against business objectives, confirming methodological soundness, and planning action. The transition to Deployment is contingent on passing the audit and re-validation gates, ensuring the model is accurate, reliable, explainable, and ethically suitable for policy use.
