# Task 1: Model Selection and Justification

## 1. Selected Model: Random Forest Regression

Random Forest was selected as the primary model for predicting continuous health outcomes in South African health datasets.

## 2. Justification

### 2.1 Data Compatibility

* **Dataset size:** 609 records with 11 features is optimal for ensemble methods.
* Random Forest handles moderate-sized datasets effectively without overfitting risks common in deep learning approaches.
* The ensemble approach provides stable predictions even with limited data.

### 2.2 Feature Handling

* **Mixed data types:** Dataset contains both categorical (health indicator types, provinces) and numerical features (sample sizes, scaled values).
* Random Forest natively handles mixed feature types without requiring extensive preprocessing.
* No assumptions about feature distributions or linear relationships required.

### 2.3 Interpretability

* **Variable importance metrics:** Random Forest provides built-in feature importance scores (%IncMSE, IncNodePurity).
* Critical for health policy contexts where stakeholders need to understand which indicators drive outcomes.
* Enables evidence-based resource allocation and intervention prioritization.

### 2.4 Robustness

* **No distribution assumptions:** Unlike linear regression, Random Forest is non-parametric.
* **Handles missing data:** Can maintain accuracy even with incomplete records.
* **Out-of-bag validation:** Internal cross-validation mechanism provides unbiased error estimates without requiring separate validation sets.
* **Resistant to outliers:** Ensemble voting reduces influence of extreme values common in health surveys.

## 3. Technical Assumptions Validation

### 3.1 Sample Size Adequacy

* 609 records split into 75% training (457 records) provides sufficient data for tree construction.
* Each tree in the forest uses bootstrap samples, increasing effective training diversity.

### 3.2 Feature Relevance

* 11 features provide adequate dimensionality for tree splits without causing high-dimensional issues.
* Default mtry parameter (sqrt of features) ensures sufficient feature diversity across trees.

### 3.3 Target Variable Properties

* Continuous target (value\_log\_scaled) is appropriate for regression Random Forest.
* Log-scaling addresses potential skewness in health outcome distributions.

## 4. Why Random Forest Over Alternatives

### 4.1 vs. Multiple Linear Regression

* Linear regression assumes linear relationships and normally distributed residuals.
* Health outcomes often exhibit non-linear patterns (e.g., threshold effects in immunization coverage).
* Random Forest captures complex interactions without manual feature engineering.

### 4.2 vs. XGBoost

* XGBoost requires careful hyperparameter tuning and is more prone to overfitting on small datasets.
* Random Forest provides more stable results with default parameters.
* Lower computational complexity for this dataset size.

### 4.3 vs. Neural Networks

* Neural networks require larger datasets (typically 10,000+ records) for reliable performance.
* Lack of interpretability conflicts with health policy requirements.
* Random Forest provides better performance with limited data.

## 5. Supporting Models

While Random Forest was selected as the primary model, two supporting models were implemented for comparison and deployment fallback:

* **Multiple Linear Regression:** Baseline model providing interpretable coefficients and deployment simplicity.
* **XGBoost:** High-performance alternative for scenarios requiring maximum predictive accuracy.

These models were built but are not documented in milestone tasks, serving as background validation of the Random Forest selection.

## 6. Summary

Random Forest Regression was chosen because it: 1. Handles the 609-record dataset effectively without overfitting 2. Manages mixed categorical/numerical features without extensive preprocessing 3. Provides interpretable variable importance for health policy decisions 4. Makes no restrictive assumptions about data distributions 5. Offers built-in validation through OOB error estimation

This selection aligns with the project’s dual objectives: achieving strong predictive performance while maintaining interpretability for health domain stakeholders.