**Alternative Approach of using Regression Model for weights of Parameters**

We can use a regression model with a dependent variable like average monthly expenditure on Information and Communication Technology (ICT) to derive feature importance scores, which can then be used as weights for the Digital Accessibility Index (DAI). This is a data-driven approach that leverages the relationship between ICT expenditure (as a proxy for digital "richness") and the independent variables (e.g., access to internet, mobile phone with internet, internet banking). The feature importance scores from the regression model can reflect how much each parameter contributes to ICT expenditure, making them a reasonable basis for weights.

**Step-by-Step Process to Use Regression Model for Weight Derivation**

**Step 1: Define the Problem and Variables**

**Objective:** To use regression to model average monthly ICT expenditure (dependent variable) as a function of digital access parameters (independent variables), then extract feature importance as weights for the DAI.

Variables:

- **Dependent Variable (Y):** Average monthly expenditure on ICT per household or individual in a village (e.g., in INR). This could include spending on mobile data, internet subscriptions, digital devices, etc.

- **Independent Variables (X):**

- Access to internet (binary: 1 = Yes, 0 = No, or % of households)

- Mobile phone with internet access (% of households or individuals)

- Access to internet banking (% of users or availability)

- (Optional) Other parameters like broadband availability, digital literacy, etc., if available.

**Assumption:**

- ICT expenditure is a reasonable proxy for digital accessibility "richness."

- Data on ICT expenditure is available for at least a subset of villages to train the model.

**Technology:** Python (Pandas for data handling, Scikit-learn for regression

**Step 2: Prepare the Dataset**

**Process:**

1. Collect Data: Gather village-level data including ICT expenditure and the independent variables. If ICT expenditure isn’t available for all 600,000 villages, we will need a representative sample (e.g., 50,000 – 200,000 villages).

2. Clean Data:

- Handle missing values in ICT expenditure or independent variables (e.g., impute using mean, median, or predictive models like k-NN).

- Standardize independent variables to a common scale (e.g., 0–1) to ensure comparability in the regression model.

- Remove outliers (e.g., using Z-scores or IQR) if they skew the results.

3. Split Data: Divide the dataset into training (e.g., 80%) and testing (20%) sets to validate the model.

**Methodology:**

- Data imputation: k-Nearest Neighbors or regression-based imputation.

- Normalization: Min-Max scaling or Z-score standardization.

**Step 3: Choose and Train a Regression Model**

**Model Selection:**

- **Linear Regression:** Simple and interpretable; assumes a linear relationship between ICT expenditure and predictors. Use coefficients as weights.

- **Random Forest Regression:** Non-linear, handles interactions between variables well; provides feature importance scores.

- **Gradient Boosting (e.g., XGBoost, LightGBM):** Robust to noisy data and missing values; offers feature importance based on gain or split frequency.

**Recommendation:** We will start with Random Forest or Gradient Boosting because:

- They don’t assume linearity.

- They provide feature importance directly, which aligns with our goal of deriving weights.

- They handle missing data and complex relationships better than linear regression.

Training Process:

1. Fit the model on the training data with ICT expenditure as the target (Y) and digital access parameters as features (X).

2. Tune hyperparameters (e.g., number of trees, max depth for Random Forest) using cross-validation to optimize performance (e.g., minimize RMSE or maximize R²).

Example (Random Forest):

- Input: X = [Internet Access, Mobile Internet, Internet Banking], Y = ICT Expenditure

- Output: A trained model predicting ICT expenditure.

**Step 4: Extract Feature Importance as Weights**

Process:

1. Obtain Importance Scores:

- For Linear Regression: Use the absolute value of standardized coefficients (β). Standardize coefficients by scaling inputs to ensure comparability.

- For Random Forest: Use the built-in feature importance (based on how much each feature reduces impurity across trees).

- For Gradient Boosting: Use feature importance based on gain (how much each feature contributes to reducing error).

2. Normalize Weights:

- Sum the raw importance scores.

- Divide each score by the total to get weights summing to 1 (e.g., if scores are [0.5, 0.3, 0.2], weights become [0.5/1 = 0.5, 0.3/1 = 0.3, 0.2/1 = 0.2]).

3. Validate Weights:

- We can check if the weights align with intuition (e.g., internet access is likely to have higher importance than banking).

- Test the model on the holdout set to ensure it generalizes well (e.g., R² > 0.6 indicates decent fit).

Example Output (Random Forest):

- Feature Importance: Internet Access = 0.55, Mobile Internet = 0.35, Internet Banking = 0.10

- Normalized Weights: [0.55, 0.35, 0.10]

**Step 5: Compute the Digital Accessibility Index (DAI)**

Formula:

- DAI = (w₁ × Internet Access) + (w₂ × Mobile Internet) + (w₃ × Internet Banking)

- Where w₁, w₂, w₃ are the weights derived from feature importance.

Process:

1. Apply the weights to the standardized values of each parameter for all villages.

2. Scale the DAI to a 0–100 range for interpretability:

- DAI\_scaled = (DAI\_raw - min(DAI\_raw)) / (max(DAI\_raw) - min(DAI\_raw)) × 100

3. Handle villages with missing ICT expenditure:

- Use the trained model to predict ICT expenditure based on available features, then compute DAI.

- Alternatively, apply weights directly to available features without prediction.

Example:

- Village A: Internet Access = 1, Mobile Internet = 0.8, Internet Banking = 0.5

- Weights: [0.55, 0.35, 0.10]

- DAI\_raw = (0.55 × 1) + (0.35 × 0.8) + (0.10 × 0.5) = 0.55 + 0.28 + 0.05 = 0.88

- DAI\_scaled = (e.g.) 88 if scaled to 0–100.

**Step 6: Validate and Refine**

Validation:

1. Model Performance: Evaluate the regression model on the test set (e.g., RMSE, R²). A good fit ensures reliable weight.

2. Index Reasonableness: Compare DAI scores with external benchmarks (e.g., government digital inclusion reports) or a sample of ground-truthed villages.

3. Sensitivity Analysis: Perturb the input data slightly (e.g., ±10%) to see if weights and DAI remain stable.

Refinement:

- If weights seem counterintuitive (e.g., banking > internet access), we will revisit feature engineering (e.g., add interaction terms) or try a different model.

- If model fit is poor (e.g., R² < 0.5), we may consider inclusion of additional predictors (e.g., income, education).