### Software Architecture for Thyroid Disease Prediction Using Machine Learning

**1. Data Collection:**

* **Source:** Hypothyroid dataset in CSV format.
* **Tools:** Pandas for data loading and initial exploration.

**2. Data Preprocessing:**

* **Handling Missing Values:** Replacing '?' with NaN and using KNNImputer for imputation.
* **Encoding Categorical Variables:** Mapping binary variables to 0 and 1, and using one-hot encoding for multi-class variables.
* **Feature Scaling:** Applying log transformation to skewed continuous features.

**3. Data Balancing:**

* **Technique:** Using RandomOverSampler from imbalanced-learn to balance the dataset.

**4. Model Building:**

* **Algorithms:** Logistic Regression, SVM, KNN, Decision Tree, Random Forest.
* **Training:** Splitting data into training and testing sets using train\_test\_split.
* **Evaluation:** Using metrics like accuracy, confusion matrix, precision, recall, and F1-score.

**5. Model Evaluation:**

* **Best Model:** Random Forest with highest accuracy and balanced performance across all metrics.
* **Cross-Validation:** Using cross\_val\_score for model validation.

**6. Model Deployment:**

* **Saving the Model:** Using pickle to save the trained Random Forest model.
* **Integration:** Deploying the model in a web application for user interaction.

**7. Tools and Technologies:**

* **Programming Language:** Python
* **Libraries:** Pandas, NumPy, Scikit-learn, Imbalanced-learn, Matplotlib, Seaborn
* **Environment:** Jupyter Notebook

**Architecture Diagram:**

+-------------------+ +-------------------+ +-------------------+

| | | | | |

| Data Collection +------>+ Data Preprocessing+------>+ Data Balancing |

| | | | | |

+-------------------+ +-------------------+ +-------------------+

|

v

+-------------------+ +-------------------+ +-------------------+

| | | | | |

| Model Building +------>+ Model Evaluation +------>+ Model Deployment |

| | | | | |

+-------------------+ +-------------------+ +-------------------+