### Low-Level Document (LLD) for Thyroid Disease Prediction Using Machine Learning

#### \*\*1. Introduction\*\*

* **Project Title**: Thyroid Disease Prediction Using Machine Learning
* **Domain**: Healthcare
* **Tech Stack**: Python, Machine Learning

#### \*\*2. Data Collection and Preprocessing\*\*

* **Dataset**: The dataset contains 3772 entries with 30 columns, including features like age, sex, TSH, T3, TT4, T4U, FTI, and various binary indicators (e.g., on\_thyroxine, sick, pregnant).
* **Libraries Used**:
  + pandas for data manipulation
  + numpy for numerical operations
  + seaborn and matplotlib for data visualization
  + scikit-learn for machine learning algorithms
  + imbalanced-learn for handling imbalanced datasets

#### \*\*3. Data Cleaning and Transformation\*\*

* **Handling Missing Values**:
  + Replaced '?' with NaN
  + Dropped columns with all missing values (e.g., TBG)
  + Used KNNImputer to impute missing values
* **Encoding Categorical Variables**:
  + Mapped binary categorical variables to 0 and 1
  + Used one-hot encoding for multi-class categorical variables (e.g., referral\_source)
* **Feature Scaling**: Applied log transformation to skewed continuous features (age, TSH, T3, TT4, T4U, FTI)

#### \*\*4. Data Balancing\*\*

* **Techniques Used**:
  + RandomOverSampler from imbalanced-learn to balance the dataset
  + Resulting in 13924 samples with balanced classes

#### \*\*5. Model Building\*\*

* **Algorithms Implemented**:
  + **Logistic Regression**
  + **Support Vector Machine (SVM)**
  + **K-Nearest Neighbors (KNN)**
  + **Decision Tree**
  + **Random Forest**
* **Evaluation Metrics**: Accuracy, Confusion Matrix, Precision, Recall, F1-Score

#### \*\*6. Model Evaluation\*\*

* **Results**:
  + **Logistic Regression**: Accuracy - 77.8%
  + **SVM**: Accuracy - 81.1%
  + **KNN**: Accuracy - 95.5%
  + **Decision Tree**: Accuracy - 98.6%
  + **Random Forest**: Accuracy - 99.2%
* **Best Model**: Random Forest with highest accuracy and balanced performance across all metrics

#### \*\*7. Model Deployment\*\*

* **Saving the Model**:
  + Used pickle to save the trained Random Forest model
  + Model file: thyroid\_model.pkl

#### \*\*8. Conclusion\*\*

* **Summary**: The Random Forest model was identified as the best-performing model for predicting thyroid disease, achieving an accuracy of 99.2%.
* **Future Work**: Potential improvements include exploring more advanced algorithms, feature engineering, and real-time model deployment.