### Detailed Project Report (DPR) for Thyroid Disease Prediction Using Machine Learning

**1. Introduction:**

* **Project Title:** Thyroid Disease Prediction Using Machine Learning
* **Objective:** To develop a machine learning model that accurately predicts thyroid disease based on patient data.
* **Domain:** Healthcare
* **Tech Stack:** Python, Machine Learning

**2. Data Collection:**

* **Dataset:** Hypothyroid dataset containing 3772 entries with 30 columns.
* **Source:** CSV file named hypothyroid.csv.

**3. Data Preprocessing:**

* **Handling Missing Values:** Replaced '?' with NaN and used KNNImputer for imputation.
* **Encoding Categorical Variables:** Mapped binary variables to 0 and 1, and used one-hot encoding for multi-class variables.
* **Feature Scaling:** Applied log transformation to skewed continuous features (age, TSH, T3, TT4, T4U, FTI).

**4. Data Balancing:**

* **Technique:** Used RandomOverSampler from imbalanced-learn to balance the dataset.
* **Result:** Balanced dataset with 13924 samples.

**5. Model Building:**

* **Algorithms Implemented:**
  + Logistic Regression
  + Support Vector Machine (SVM)
  + K-Nearest Neighbors (KNN)
  + Decision Tree
  + Random Forest
* **Training and Testing:** Split data into training and testing sets using train\_test\_split.

**6. Model Evaluation:**

* **Metrics:** Accuracy, Confusion Matrix, Precision, Recall, F1-Score.
* **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
* **Integration:** Deployed the model in a web application for user interaction.

**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.

**Appendix:**

* **Code Snippets:**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

import pickle

# Load data

data = pd.read\_csv('hypothyroid.csv')

# Data preprocessing steps

# ...

# Model building

X = data.drop('Class', axis=1)

y = data['Class']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = RandomForestClassifier()

model.fit(X\_train, y\_train)

# Save the model

with open('thyroid\_model.pkl', 'wb') as file:

pickle.dump(model, file)