Thyroid disease prediction, classification and early detection

Aryan Malhotra - 2019ABPS0893P

Problem statement and research goals

Thyroid disease affects millions of people worldwide and thus represents a significant global health concern.

This project aims to leverage a labeled dataset containing parameters on general health as well as relevant biomarkers for Thyroid (TSH, T3, TT4 etc.) to create a machine learning model. This model aims to predict, classify and facilitate early detection of Thyroid disease and its types

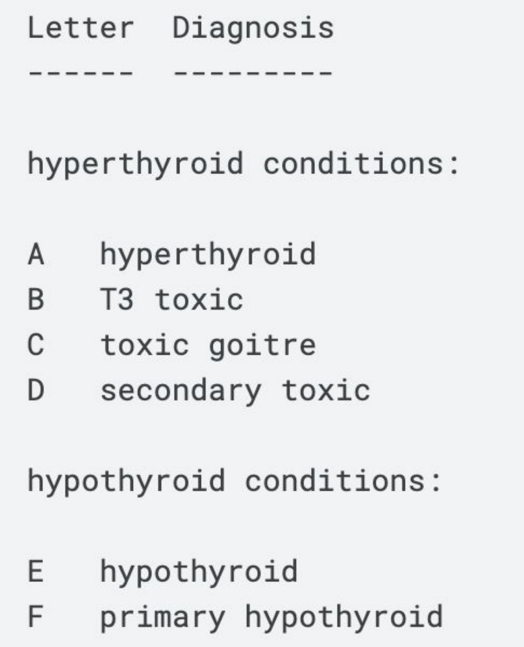
Dataset

- Garvan institute of medical research, Australia

- **28** input parameters (boolean and numeric type) across general health and thyroid relevant biomarkers

- Data shape is a table with 9173 rows * 31 columns

- Target labels are alphabets ranging from A to T, classifying specific types of thyroid disease as well as other medically significant observations



```
Letter  Diagnosis
------  ---------

hyperthyroid conditions:

A    hyperthyroid
B    T3 toxic
C    toxic goitre
D    secondary toxic

hypothyroid conditions:

E    hypothyroid
F    primary hypothyroid
```

Random forest classification algorithm

Random forest algorithms combine multiple decision trees to make predictions. Two specific reasons for choosing this algorithm over others is

- **Non linearity and feature importance** : Handling non linear relationships in data, Identification of feature importance

- **Ensemble learning** - Utilising multiple decision trees reduces overfitting in the model, Outliers in data can be handled effectively

Preprocessing and hyperparameter tuning

Numpy, **Pandas** & **Scikit-learn** python libraries will be used for data pre processing

GridsearchCV - To identify best hyper parameters and perform cross validation

SMOTE - To address class imbalances, Smote will generate synthetic data for minority target classes to ensure that the model does not have a bias.

Metrics from the model as well as any other relevant methods to boost model performance will be utilised

Time and Space complexity analysis

For the algorithm, computation time will be required during splitting process of each tree.

During training, T(o) will be proportional to number of trees, no of features, no of samples and depth of tree.

During predicting, **T = o(n_trees*max_depth)** is the maximum required time

Space complexity for the dataset will be proportional to **X = (n_features * n_samples)**.

The maximum space complexity for the decision trees will be proportional to, but less than **(n_trees*X)**

Methodology

The project will be executed using Python and maintained on a Jupyter notebook.

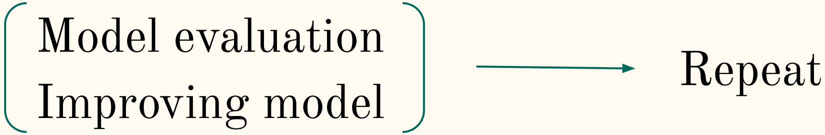
The broad steps are as follows:

Data collection & Preprocessing

Exploratory data analysis - Matplotlib & Seaborn

Hyper parameter tuning

Model training



Model evaluation

Improving model

Repeat

Final insights, results, documentation & submission

Expected Results

The primary goal of the project is to classify patients as having thyroid disease or not.

The model aims to predict further sub classifications such as type of thyroid disease (hyper, hypo and its sub types)

The project also aims to identify patients at risk of contracting thyroid disease for early detection and intervention