Thyroid disease prediction, classification and early detection

Aryan Malhotra - 2019ABPS0893P

Index

1 - Problem statement

Thyroid disease affects millions of people worldwide and thus represents a significant global health concern. The timely diagnosis and correct classification of thyroid is extremely important, given the wide range of conditions (from hypothyroid to hyperthyroid). The need for a reliable thyroid disease prediction system is increased when we realize that there is an even wider variety of treatment options (antithyroid drugs, replacement therapy, surgery etc.). The identification of the appropriate diagnosis reaching each individual patient is of paramount importance.
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 not only classify thyroid diseases, but also facilitate early detection and intervention.

2 - Benchmarked Thyroid dataset

The dataset that I am using for my analysis is a dataset taken from the Garvan institute of Medical research in Australia. This dataset has also been uploaded to the UC Irvine machine learning repository website.

The dataset uses 28 input parameters ranging from general information and health to presence and levels of certain biomarkers relevant with thyroid disease. The input data types include boolean data as well as continuous values (numeric type) for indicating biomarker levels.

The shape of the data is a table with 9173 rows and 31 columns.

3 - Algorithms to be used

The algorithm using which the model will be trained will be the **Random forest classification algorithm**. The reason for choosing this algorithm over other popular classification algorithms like SVM or logistic regression is twofold -

Non linearity and feature importance : Random forests algorithms are capable of handling non linear relationships (such as in the case of biomarker levels and disease classification).

Random forests also help us identify feature importance which can help in understanding of the disease, its causes and early detection.

Ensemble learning - One of the biggest advantages of this algorithm is that it utilizes multiple decision trees to make a prediction. This reduces overfitting in the model and will handle any outlier values effectively, without compromising model performance.

One of the disadvantages of using this algorithm is the lack of interpretability in the prediction process of the algorithm due to the large number of individual trees being used.

GridsearchCV will be utilized to identify the best hyper parameters for training the model and performing cross validation.

SMOTE will be utilized after training the first version of the model to address class imbalances. Smote will help in generation of synthetic data for minority classes to ensure that the model does not have a bias towards the majority class

Other methods to boost performance of the algorithm will also be deployed as and when required

4 - Methodology and data analysis

The language for the ML model will be scripted in python for its advanced libraries capable of carrying out ML tasks. All code will be written and maintained on a Jupyter notebook for easy visualization of data and segmenting different parts of the project.

Data collection, preprocessing, EDA, hyper parameter tuning, model training and evaluation followed by continuous improvements will be the broad method for carrying out the project.

The initial data analysis will focus on summarizing statistics and analyzing correlations in biomarker levels and thyroid disease. Post model training, analysis such as feature importance and pairwise comparisons (thyroid positive and negative groups) of biomarker levels will be carried out. F1 scores and other metrics will be used for feature engineering and improving model performance.

Python libraries such as matplotlib and seaborn will be utilized for data visualization purposes.

5 - Expected results

The primary goal of the project is to classify patients as having thyroid disease or not. Further sub classifications such as type of thyroid disease will also be predicted. The project also aims to identify patients at risk of contracting thyroid disease (through biomarker levels analysis and correlations).

The true strengths of the model will be unleashed when the model is deployed in a clinical setting. In such a setting, it will not only work on classification of the disease, but will also work on reducing false positives and false negatives during diagnosis of patients.