Diagnosis of Thyroid Diseases from Blood Work Results

1 Problem Statement

I would like to tackle the problem of predicting whether and what type of thyroid disease a patient has judging from their blood work data. This problem is interesting because one can diagnose a patient between 20 different diseases just from looking at the same data.

2 Methodology

2.1 Input

The blood work and thyroid disease status dataset is collected from an open-source database[1]. The data is presented as a 9712×31 matrix of numbers, strings, and Booleans which includes the patients' current medications, hormone levels in blood, and the doctors' diagnoses.

2.2 Processing

Since there are 20 different diagnoses in the dataset, the diagnoses will first be encoded with one-hot encoding; notably this also allows the model to diagnose the patient with multiple diseases since the doctors also did so. The dataset will then be split into training and testing data with an 80:20 ratio. It will then be fed into an AI model (provisionally, random forest classifier). The AI model will then be trained to predict the diseases of the patients. The final AI model will be chosen based on its accuracy score.

2.3 Output

The output of the model will be a one-hot encoded vector for different diseases as the model's prediction of the diagnoses.

3 Expected Results

The expected results would be ~85% accuracy compared to the ground truth. The accuracy would be measured by getting the accuracy scores from the model's library. The results can also be manually evaluated using a confusion matrix.

[1] Quinlan, R. (1986). Thyroid Disease [Dataset]. UCI Machine Learning Repository. doi:10.24432/C5D010