Understanding Hypothyroidism: An Analytical Approach

Final Project Report
Subash Matu
Data Science Initiative
https://github.com/Subashmatu08/Data1030FinalProject

1. Introduction

Reasoning

Thyroid related problems are growing in number now more than ever. Several articles claim it to be the most common disorder. The two main type of thyroid related problems are Hyperthyroid and Hypothyroid. In this analysis let's explore Hypothyroid and why it occurs. Every 5% of the general population suffer from Hypothyroidism in which 5% aren't properly diagnosed.

According to existing research there are three main hormones that deal with thyroid, they are T3, T4 and TSH. T3 and T4 being released by the Thyroid gland and TSH released by the pituitary gland. TSH decides how much of the other hormones to be released. The pituitary gland releases TSH (Thyroid Stimulating Hormone) which goes and acts on the thyroid gland to secrete T3 and T4 which help the body maintain in metabolic rate, temperature of the body, body weight, etc.

Data is collected from the CI Machine Learning Directory through Kaggle.

Importance

Early diagnosis by screening and understanding the factors can lead to better treatment and preventive measures as this is a prevalent disorder that impacts a significant portion of the population.

Current Research

There have been several approaches to make a Screening Test for Hypothyroid. A few of them can be found on Kaggle as well. Those approaches show accuracies of 99.5% and 99.13% which suggests that the models have been trained using their best algorithms. The most commonly used algorithms found for this dataset were XGBoost.

2. EDA

The general shape and size of the data set:
The dataset contains 3622 entries, with 28 attributes.

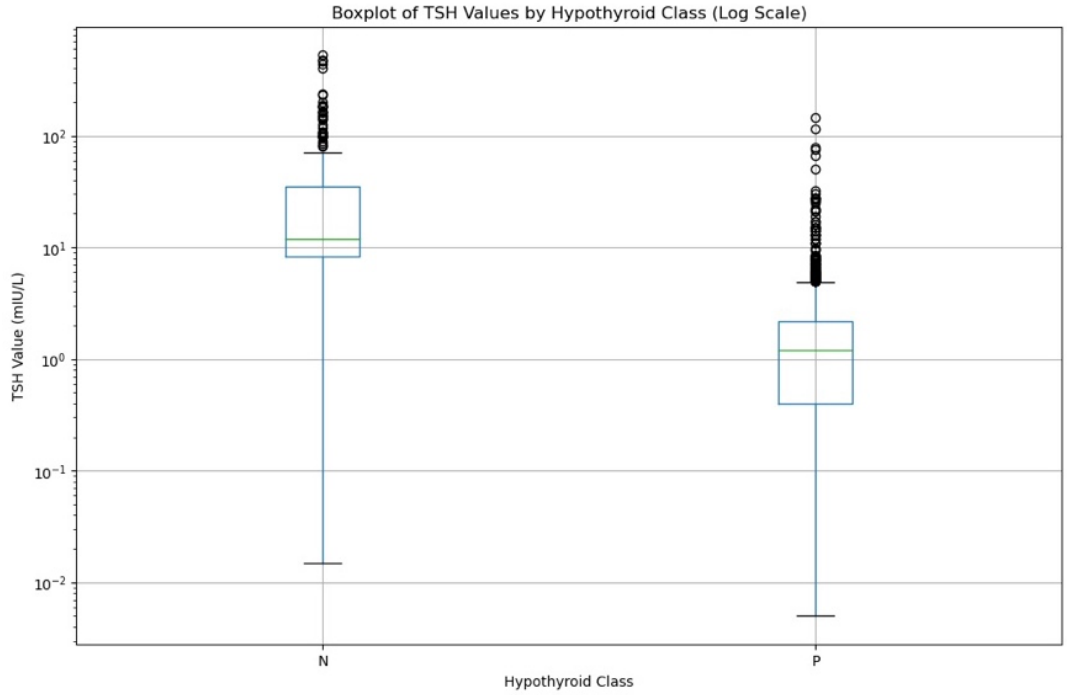

Figure 1. Boxplot of TSH Values by Hypothyroid

In regards to TSH levels and their connection to hypothyroidism, the 'Positive' group (presumed to have the illness) displayed a higher TSH level median than the 'Negative' group. The correlation points towards hypothyroidism being linked to heightened TSH levels. Patients with hypothyroidism belonging to the 'P' group have a wider interquartile range (IQR) compared to the 'N' group, indicating that the variation in TSH levels is greater. This can be observed from the height of the boxes.

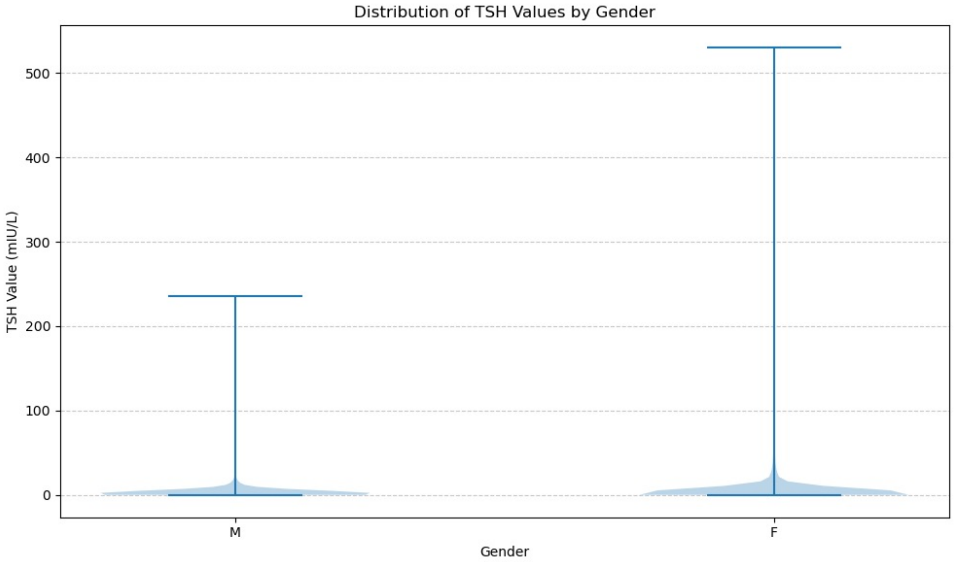

Figure 2. Distribution of TSH values by Gender

Two clusters of data points are displayed in the plot: one for males and one for females. We cannot form a solid conclusion but from this graph we can say that there are more females reported with higher TSH values than males.

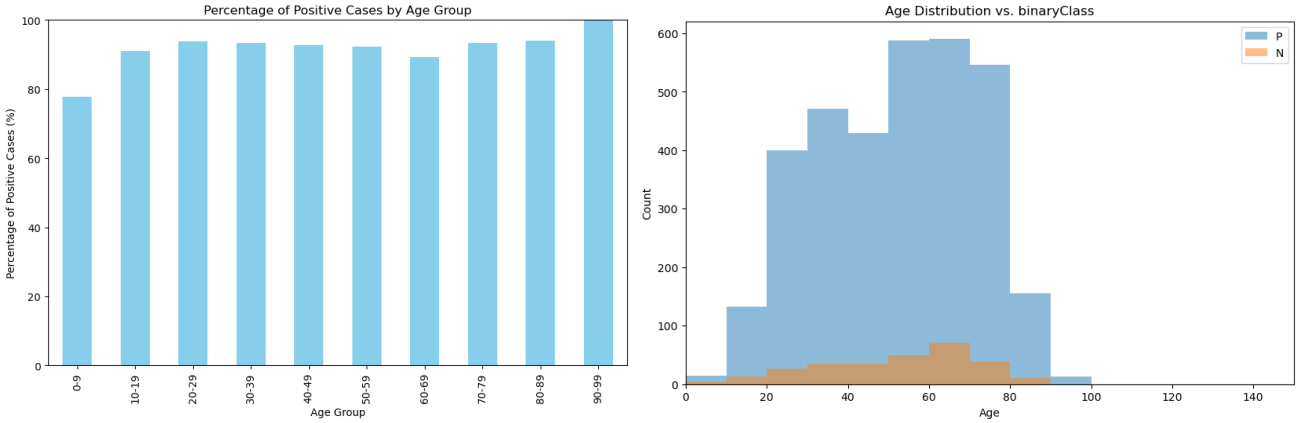

Figure 3. Plots against Age and count of the positive cases

From the age group against positive cases we can observe that the number of positive cases increases with age suggesting that people of more age have higher tendency of getting such disorder.

Missing Values

The missing values in those features are :
age 0.000276
TSH 0.097211
T3 0.205744
TT4 0.059928
T4U 0.101353
FTI 0.100801
age_group 0.000276

All of the features with missing values are continuous features.

3. Methods

Dataset Split

Since it's a classification problem I chose to split it using train-test-val with ratios of 6:2:2 which ended up with the size (2172, 724, 725)

Preprocessor

For the categorical data I used OneHotEncoder and for the continuous data I used Min-Max scalar.

ML Pipeline

A total of 5 ML algorithms were used for this analysis. They were Logistic regression, KNN, SVM, XGBoost and Random Forest. For each of these several parameters were tuned with GridSearchC. Table () shows the best parameters detected for each one.

For XGBoost, missing values need not be handled as it takes care of it with in its approach while the others had to be delt with. The method of approach taken was multivariate imputator as there was no correlation among the missing values.

Logistic Regression	C:100, penalty: l2
KNN	Metric: Manhattan, n-neighbours: 5
SVM	C: 100, gamma:1
XGBoost	gamma: 0.1 max_depth: 4 n_estimators:100
Random Forest	max_features: 1.0 Max_depth: 10

Table 1. Table of various best parameters when hyperparameter tunning various ML Algorithms

The baseline for this was 0.904828. From the graph below we can observe that Random Forest had the best performance. The various Test accuracy scores for each Algorithm is as below:

Logistic Regression	0.936552 (93.65%)
KNN	0.925517 (92.55%)
SVM	0.943448 (94.34%)
XGBoost	0.955900 (95.59%)
Random Forest	0.964138 (96.41%)

With this data we can conclude that the highest accuracy was for Random Forest with test accuracy of 96.41%

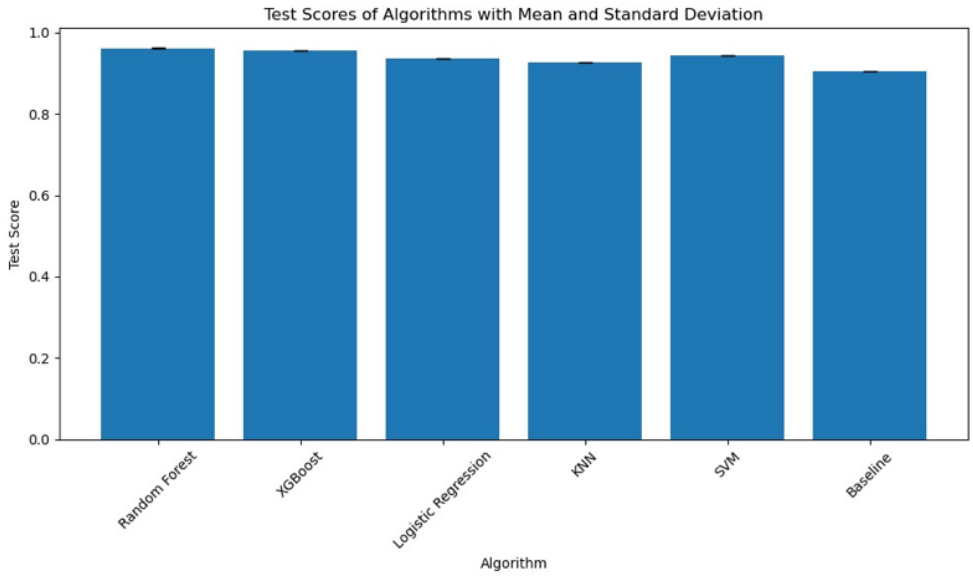

Figure 1. Bar graph showing Test scores of algorithms with Mean and Standard Deviation

The graph above shows the various algorithms and their test scores with mean and standard deviation. Suggesting which algorithm gives the best score compared to the baseline.

4. Results

With this model we aim to replace traditional screening tests. As screening test varies from diagnosis, we have the liberty to accept True Positives.
When a person is detected be True positive, they are given a medication which is generally given to Hypothyroid patients resulting in 'Factitious Hyperthyroidism'. The symptoms are although severe, it can be treated and can be detected quickly. This however cannot be said for the patients who have Hypothyroid but are given no medication as they were detected to be FN for which the effects are detrimental. The symptoms are long term and cannot be identified as fast. For this reason, our goal is to reduce as many FN as possible at the cost of having some FP. This can be achieved by Recall.

Recall: True Postive / True Positive + False Negative

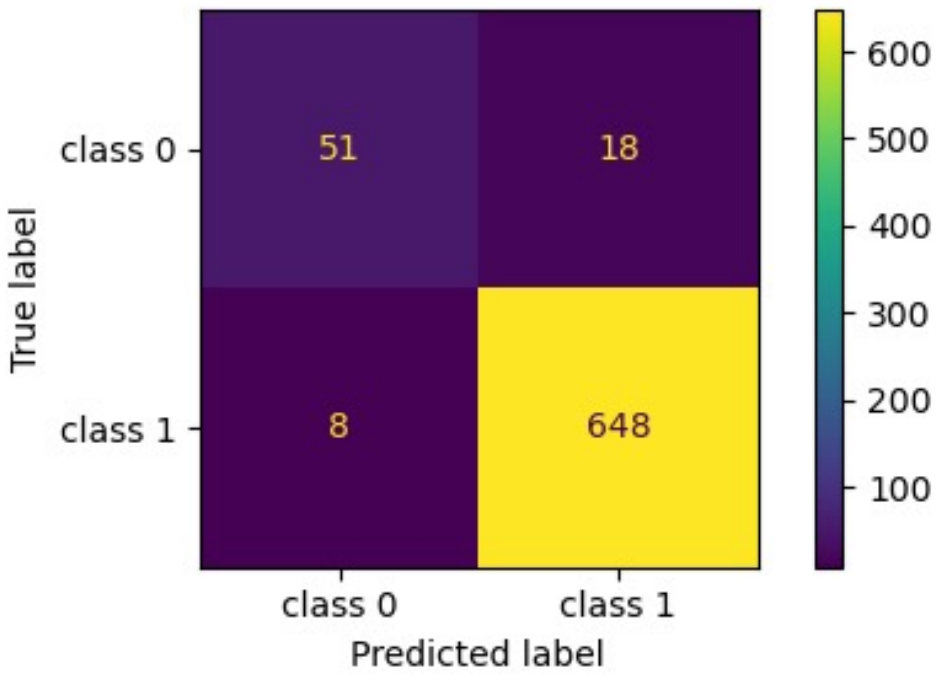

Figure 1. Confusion Matrix showing between True label and Predicted label

The confusion matrix displays how many of the model's predictions were accurate and inaccurate. The diagonal cells (51 and 648) show how many of the model's predictions came true. The off-diagonal cells (18 and 8) show how many wrong predictions the model made.

The recall value for Random Forest is 0.99

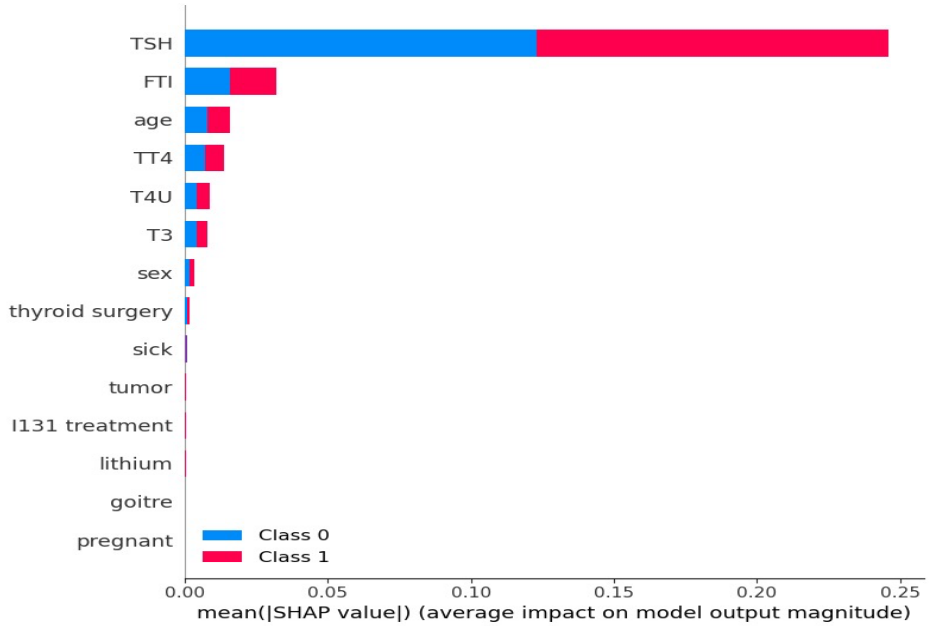

Figure 2. Summary Plot of the mean(Shap Values) and various features

The image displays the average effect of several factors on a machine learning model's output. Many factors, including TSH, FTI, age, TT4, TA4, T3, thyroid surgery, sex, tumor, I131 treatment, lithium, goitre, and pregnancy, are listed on the y-axis.

This graph is used to indicate which of the features are the most influential and possible tell the model to decide on the output during training.

From the graph we can observe with the mean SHAP Values that TSH is indeed the most influential feature followed by the FTI (Free T4 Index) and age of the patient. This is true to the research as well as TSH promotes the thyroid hormone to release T3, and T4, suggesting that it has a high say in whether the patient can be facing Hypothyroid disorder.

5. Outlook

The main weak point of my modelling approach was the choice multivariate imputation when dealing with missing values in continuous data which is not recommended when dealing with a dataset of people's health. Although it was taken into consideration when creation of this model several factors led to deciding to not use Sub Model process.

Even though there a large number of data points we can observe that there are many outliers in the data. This could be because of the data set which questions its legitimacy.

The model could be benefitted with some more insights regarding the patients as in their wight, height, general calorie intake as all these tend to base the result according to research.

6. Reference

[1] https://www.kaggle.com/code/sam4om/eda-xgbclassifier
[2] https://www.kaggle.com/datasets/yasserhessein/thyroid-disease-data-set
[3] https://www.mayo.edu/research/clinical-trials/diseases-conditions/thyroid-disease/
[4] https://en.wikipedia.org/wiki/Thyroid
[5] ollowell J.G. Staehling N.W. Flanders W.D. et al.
Serum TSH, T(4), and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III).
J Clin Endocrinol Metab. 2002; 87: 489-499