Thyroid detection and analysis

Introduction

Your thyroid produces thyroid hormone, which controls many activities in your body, including how fast you burn calories and how fast your heart beats. Diseases of the thyroid cause it to make either too much or too little of the hormone. Depending on how much or how little hormone your thyroid makes, you may often feel restless or tired, or you may lose or gain weight. Women are more likely than men to have thyroid diseases, especially right after pregnancy and after menopause.

Purpose and Motivation

The purpose of the project is to analyze and predict a person has Thyroid disease or not.

This project will help further researchers to make decisions and build models and collect more data accordingly.

Project Phases

This project is divided into four sections:

1. Data Engineering: Cleaning the dataset to fit it according to model.

2. Exploratory Data Analysis: Exploring the data and finding useful insights using bar plots to help serve the purpose of the project.

3. Managing the NaN values: Replacing, and dropping NaN values to make data ready to fit into the model.

4. Machine Learning: Creating different Machine Learning models and finding the best fit model.

Data Engineering

Typecasting Age column from string to integer.

Dropping the values with Age greater than 400.

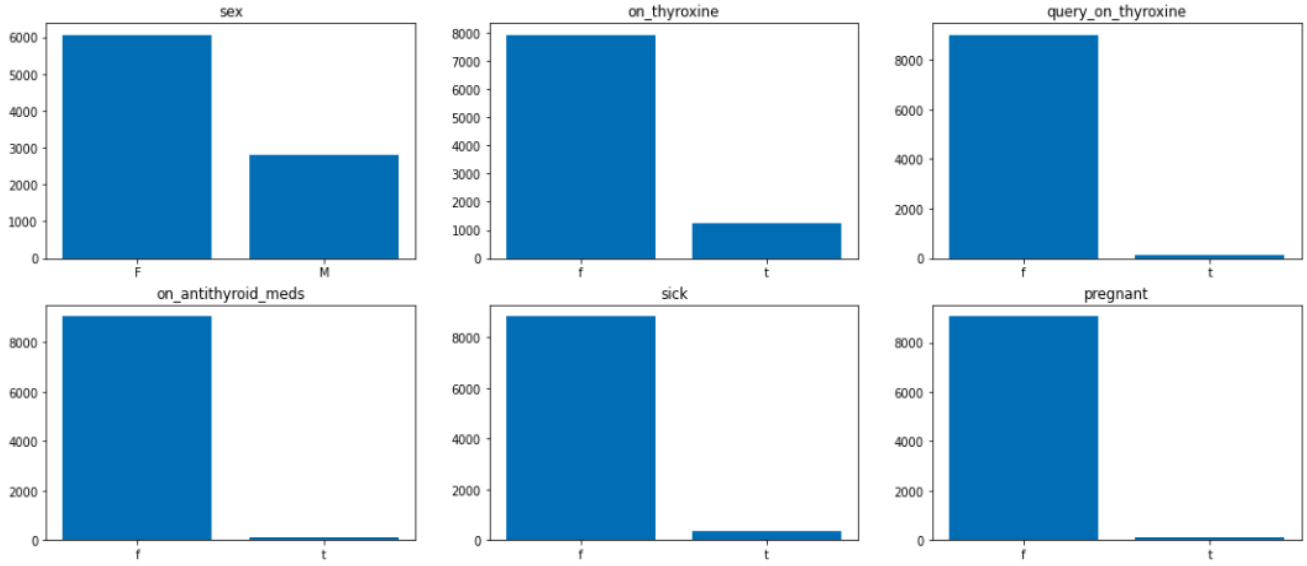
Removing prefixes from the names of dependent columns.

Dropping TBG feature which has more than 90% of the NaN values which cannot be replaced as it can add bias to the dataset.

Thyroid detection and analysis

Exploratory Data Analysis

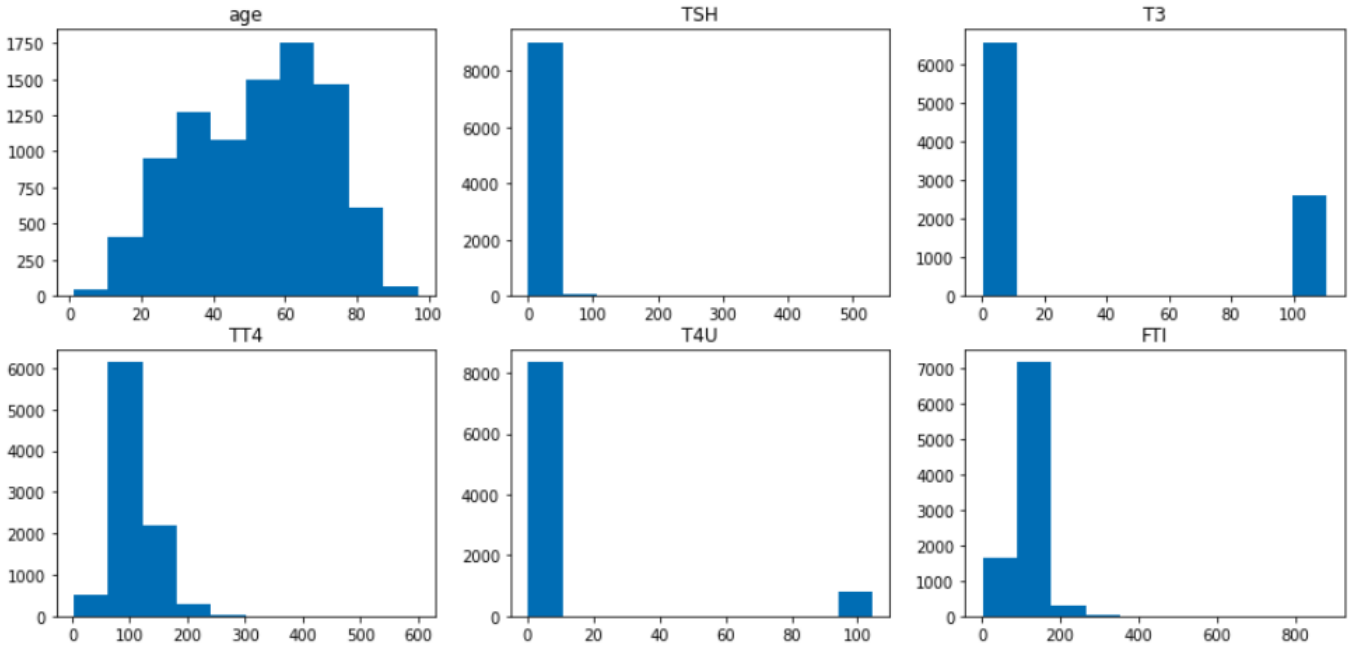
Categorical data



Observations:

We can see that the ratio of true by false value is very low in most of the categorical variables except on_thyroxine, and Sex feature. We can see that most of the cases of female patients. With this we can conclude that the research is mostly focused on female sex as they get this disease more compared to male.

Numerical data

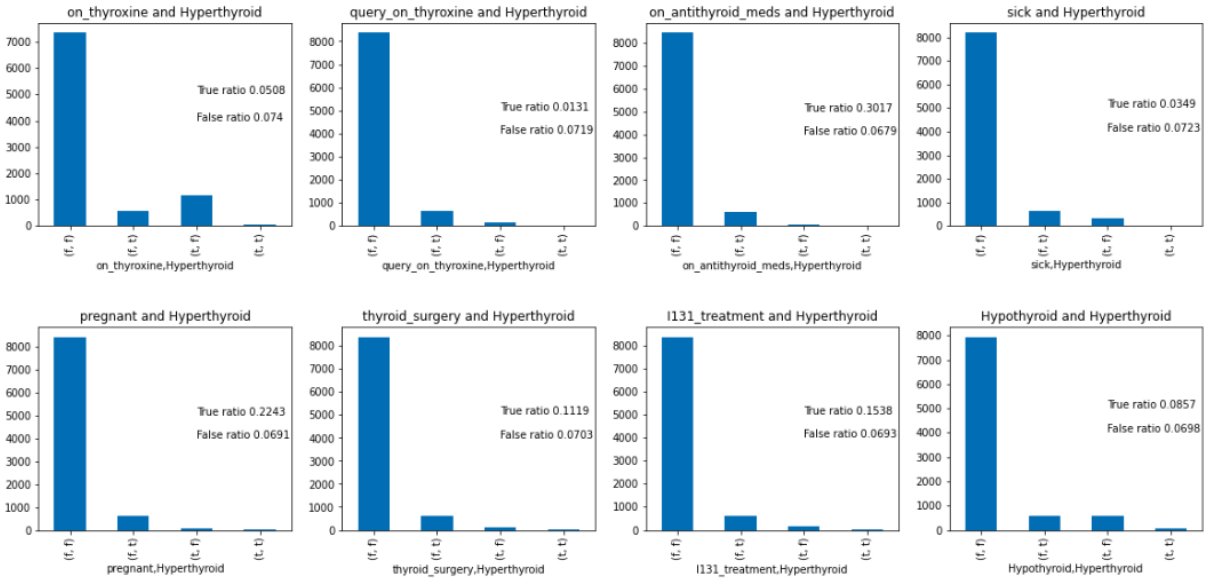


Thyroid detection and analysis

Observations:

- The plot of Age, TT4, T4U are near to normally distributed.
- TSH, T3, and FTI are rightly skewed.
- Looks like we've all the values of age in the range 1 to 97.
- People coming between the age 60-70 are more prone to this disease.

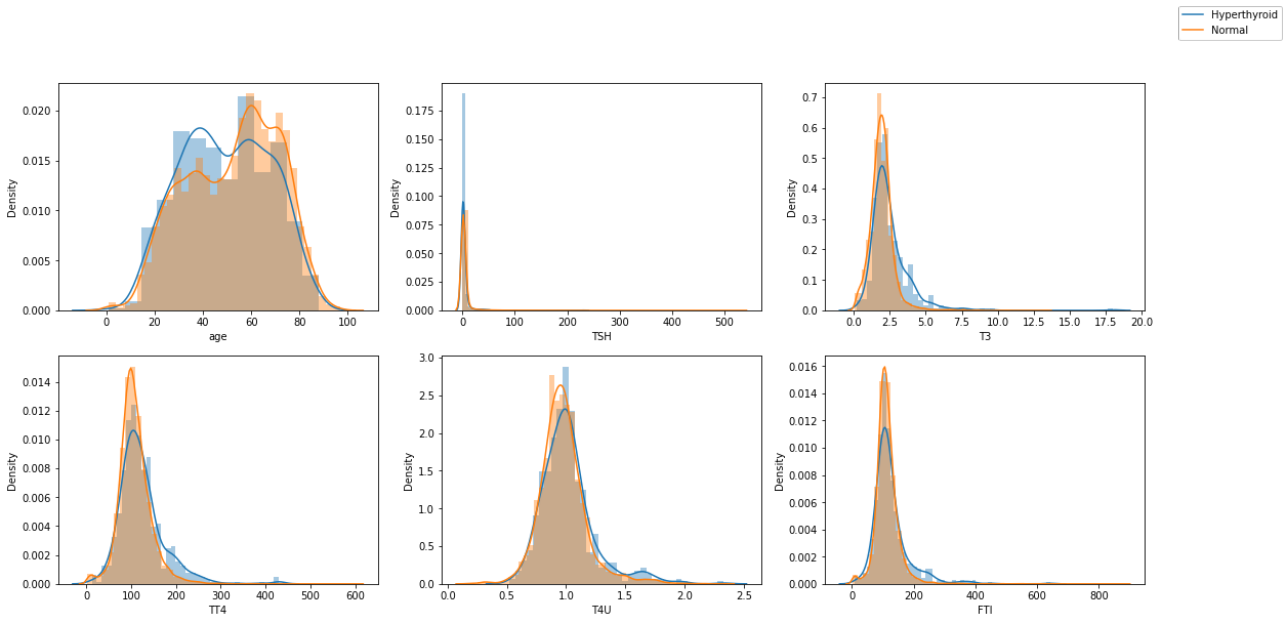
Grouping categorical features with Hyperthyroid column



Observations:

- From the above plot and calculation we can say that the women with pregnancy don't have Thyroid as compared to the women who don't.

Correlating Numeric features with Hyperthyroid

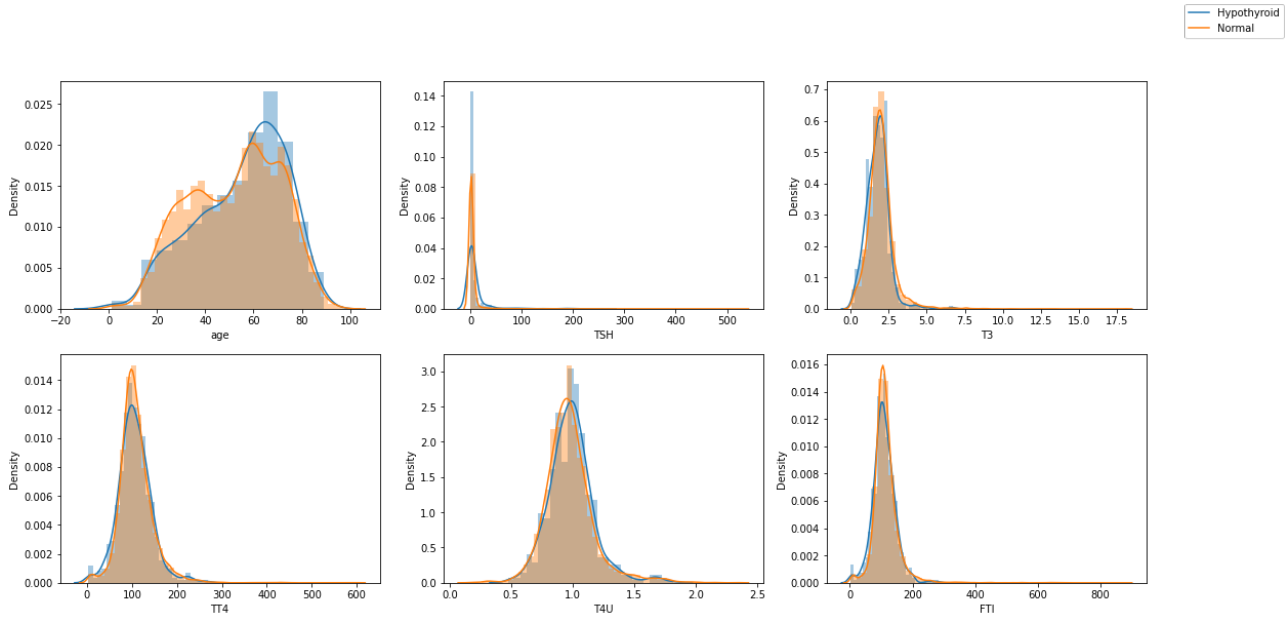


Thyroid detection and analysis

Observations:

- We've more people with no Thyroid in TSH feature.
- In age feature the people with Hyperthyroid are densly populated in the age range of 40-60 and Normal people are populated between 60-75.

Correlating Numeric features with Hypothyroid



Observations:

- In the age feature the data is more deviated in Thyroid cases compared to normal cases.
- Conversly to the age feature, in the TT4, TSH, and T3 feature the data is more has more deviated in Normal cases than the Thyroid cases.
- In T4U the data is normally distributed in both the cases.

Replacing NaN values

I’ve built an algorithm which uses feature correlation to replace the NaN values of a feature.

It chooses the highest correlated feature with that feature and calculates the modes, sum them up, finds the data equal to the sum of the modes, accesses the feature and calculated the mean which is used to replace with NaN values.

Thyroid detection and analysis

Model Building

As the dataset contains both the type of Thyroid diseases i.e Hypothyroid and Hyperthyroid I've split the dataset in two dataframes.

I've used 4 different models to predict the Thyroid disease.

Linear Regression: This model uses simple Linear Regression model to fit the data and then classify the output based on hard coded threshold which is 0.1.

Logistic Regression: This model creates a probabilistic curve with lowest value 0 and highest value as 1. It select a threshold on it's own and classifies the data point.

Kmeans clustering: This model creates clusters in higher dimensions and classifies a data point according to that.

Multi-layer perceptron classifier: This model uses neural network and is able to understand non-linear relationships between the features.

Ordinary Least Squares: This model uses Least Squares method to find a best fit line for the data. I've used Statsmodel API for the implementation.

Accuracy metric: F1 score as a metric to calculate accuracy as it the dependent variable is binary i.e True and False. The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 * (precision * recall) / (precision + recall)

Results for Hyperthyroid:

Linear Regression:

```
LinearRegression()
F1 Score: 0.2226148409893993
```

col_0	0	1
Hyperthyroid		
0	1270	380
1	60	63

Thyroid detection and analysis

Logistic Regression:

```
LogisticRegression()
F1 score: 0.04580152671755726
```

col_0	0	1
Hyperthyroid		
0	1645	5
1	120	3

Kmeans model:

```
KMeans(max_iter=5000, n_clusters=2)
F1 score: 0.11635865845311431
```

col_0	0	1
Hyperthyroid		
0	397	1253
1	38	85

MLP classifier:

```
MLPClassifier(max_iter=1000, random_state=1)
F1 score: 0.07407407407407408
```

col_0	0	1
Hyperthyroid		
0	1643	7
1	118	5

OLS classifier:

```
Ordinary Least Squares
F1 score: 0.2510121457489879
```

Hyperthyroid	0	1
row_0		
0	1557	92
1	93	31

Thyroid detection and analysis

Top 15 important features

```
on_antithyroid_meds_t    0.105423
I131_treatment_t         0.054400
pregnant_t               0.051960
goitre_f                 0.050056
T3                       0.042738
psych_f                  0.039447
TT4_measured_f           0.037033
TSH_measured_f           0.036517
query_on_thyroxine_f     0.036188
FTI_measured_t           0.035486
thyroid_surgery_t        0.030606
hypopituitary_t          0.030262
on_thyroxine_f           0.025949
T3_measured_t            0.022131
tumor t                  0.021717
```

OLS description:

OLS Regression Results

Dep. Variable:	Hyperthyroid	**R-squared:**	0.063
Model:	OLS	**Adj. R-squared:**	0.060
Method:	Least Squares	**F-statistic:**	19.09
Date:	Tue, 20 Dec 2022	**Prob (F-statistic):**	4.25e-82
Time:	18:19:52	**Log-Likelihood:**	-259.66
No. Observations:	7088	**AIC:**	571.3
Df Residuals:	7062	**BIC:**	749.8
Df Model:	25		
Covariance Type:	nonrobust		

Results for Hypothyroid:

Kmeans:

```
KMeans(max_iter=5000, n_clusters=2)
F1 score: 0.1075268817204301
```

col_0	0	1
Hypothyroid		
0	1245	405
1	93	30

Thyroid detection and analysis

Logistic regression:

```
LogisticRegression()
F1 score: 0.0
```

col_0	0
Hypothyroid	
0	1650
1	123

Linear Regression:

```
LinearRegression()
F1 score: 0.19480519480519484
```

col_0	0	1
Hypothyroid		
0	1356	294
1	78	45

MLP classifier:

```
MLPClassifier(hidden_layer_sizes=500, max_iter=10000, random_state=1)
F1 score: 0.03870967741935485
```

col_0	0	1
Hypothyroid		
0	1621	29
1	120	3

OLS model:

```
Ordinary Least Squares
F1 score: 0.13229571984435798
```

Hypothyroid	0	1
row_0		
0	1533	106
1	117	17

Thyroid detection and analysis

Top 15 important features:

FTI_measured_f	0.053302
TBG_measured_f	0.049371
T4U_measured_t	0.041663
on_thyroxine_t	0.039639
TT4_measured_f	0.031056
pregnant_f	0.026515
on_antithyroid_meds_f	0.026106
I131_treatment_t	0.023203
tumor_f	0.017111
hypopituitary_f	0.016768
sex_F	0.016457
goitre_f	0.015382
TSH_measured_t	0.014714
psych_f	0.012414
T3_measured_f	0.009671

OLS summary:

OLS Regression Results

Dep. Variable:	Hypothyroid	**R-squared:**	0.033
Model:	OLS	**Adj. R-squared:**	0.029
Method:	Least Squares	**F-statistic:**	9.511
Date:	Tue, 20 Dec 2022	**Prob (F-statistic):**	6.79e-36
Time:	19:21:25	**Log-Likelihood:**	-250.79
No. Observations:	7088	**AIC:**	553.6
Df Residuals:	7062	**BIC:**	732.1
Df Model:	25		
Covariance Type:	nonrobust		

Thyroid detection and analysis

Conclusions

- The data is mostly imbalanced towards non-thyroid cases.
- Thyroid disease is mostly found in Women.
- Thyroid disease mostly occur Women in postpartum or after menopause period.
- The dataset has many features with False values compared to True values.

Limitations and Future work

- The dataset is missing if a female has Meopause or not.
- There are lots of interesting features to explore and relate to the Thyroid feature.
- There are more combinations of feature which affects the Thyroid disease.
- More data can be obtained by scraping data from other websites and synthetically generating data using ML models.
- There are not much features which correlates with the Thyroid features.

References

Thyroid disease

Statistical Inference

Ordinary Least Squares, OLS Statsmodels

Hyper vs Hypo Thyroid