THYROID DISEASE PREDICTION MODEL- PREDICTIVE ANALYTICS (20XD43)- M.Sc Data Science 2nd year

By, Nithin.V (22PD24)

Harshan.M.V (22PD14)

Sujan.S (22PD35)

Gowtham.S (22PD13)

ABSTRACT:

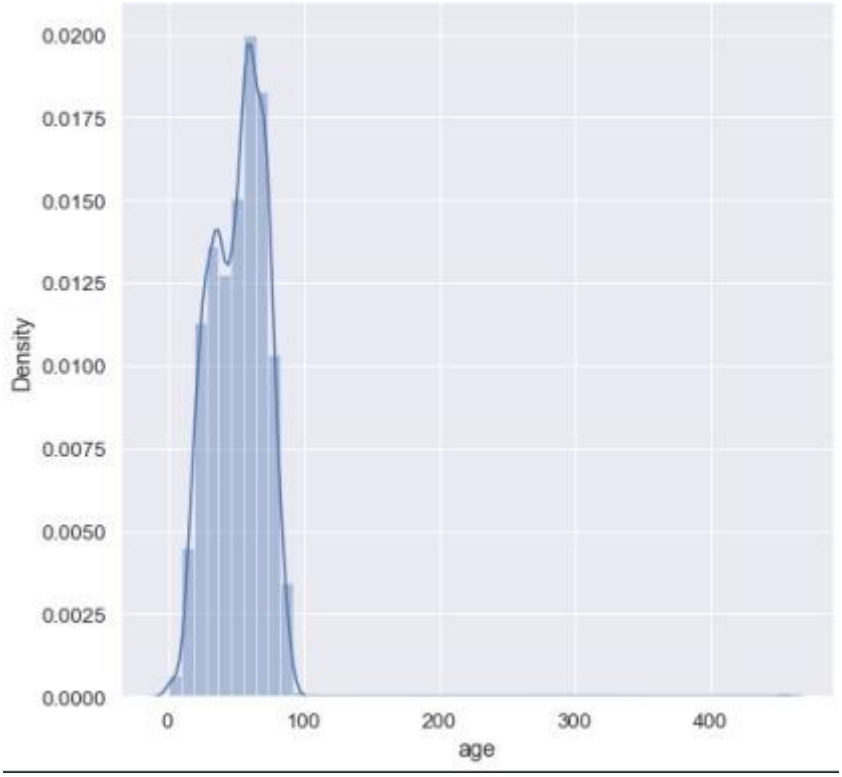
PART-1 The dataset predominantly consists of categorical data, so logistic regression will be an ideal application for the same. The Logistic Regression model is formulated both, by manual method and using the inbuilt libraries. The Entire dataset is divided into training and testing data. The first 75% of the data is taken as a testing data to build a model and we use the rest to predict it. The predicted model is then used to find the beta i's which indicates how much each factor is affecting the person having thyroid. Higher the values of Beta i's, higher is that factor influencing. Thereafter the accuracy of the model is tested and the accuracy percentage is calculated. The accuracy turns out to be 92%. Thus, we now can give the real-time data and predict if a person is having thyroid or not with 92% success rate.

```
The Beta i's vals are:   [[ 1.66393781 -0.11641126  0.43241729 -0.21088505  1.05178596  0.03344461
  -0.30673769 -0.89078957  0.39910449 -0.00825258  0.82824032 -0.33301339
   0.05837711]]
-----------------------------------------------------------------
the maximum value of the Beta is:  1.6639378111100578
Thus the factor influencing the most is:  on 1hyroxine
-----------------------------------------------------------------
The factor that has most neglib;e or least impact is:  -0.8907895726814492
Thus the attribute is:  query hypo1hyroid
intercept of the model is:  [2.3785736]
The predicted value is:  [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
The accuracy of the model is:  92.31788079470199 %
----------------------------------------------------------------------------------------------------------------
PS E:\Predictive Analysis Project>
```

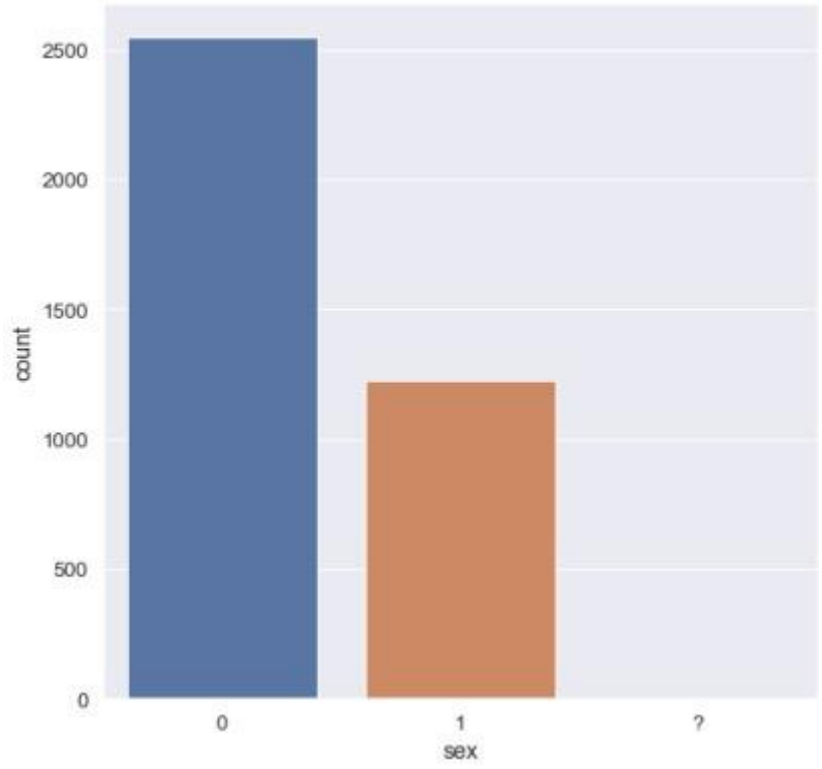
PART-2 Seaborn and Matplotlib libraries are used for data visualization and analysis. Rows containing non-numeric values in the 'age' column are filtered out, ensuring data integrity.

The script then sets aesthetic parameters for all plots, ensuring consistent styling throughout. Several types of plots are generated to explore different aspects of the dataset:

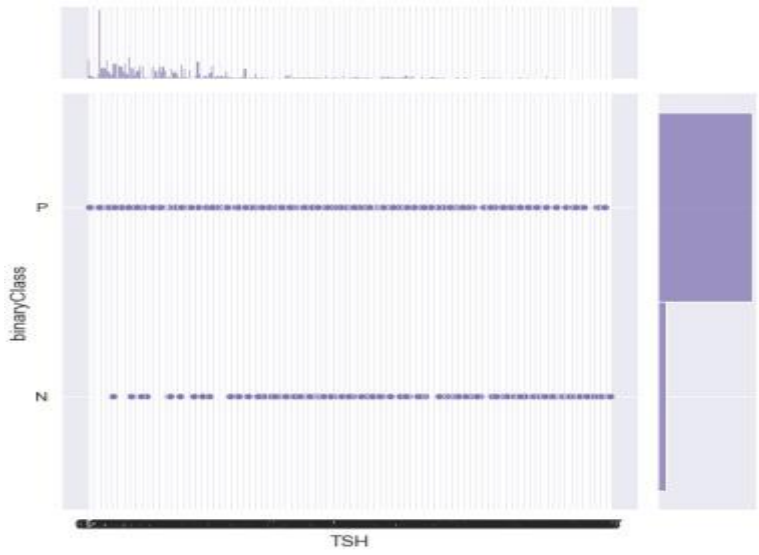
A distribution plot (histogram) is created for the 'age' column using Seaborn's distplot() function. This plot visualizes the distribution of ages in the dataset, providing insights into the age demographics.



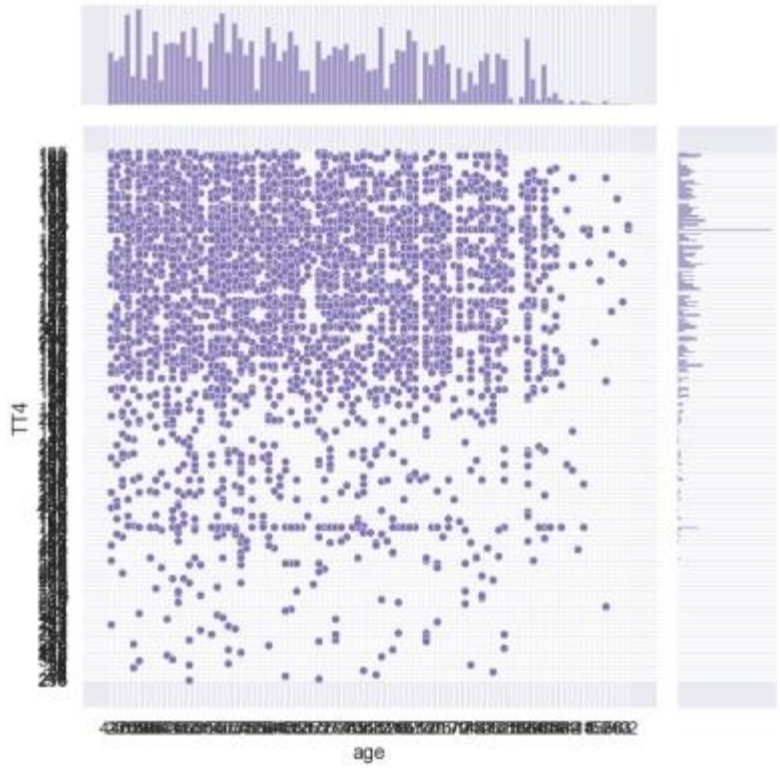
A count plot is generated for the 'sex' column using Seaborn's countplot() function. This plot displays the frequency of each category (e.g., male and female) in the 'sex' column, allowing for gender-based analysis.



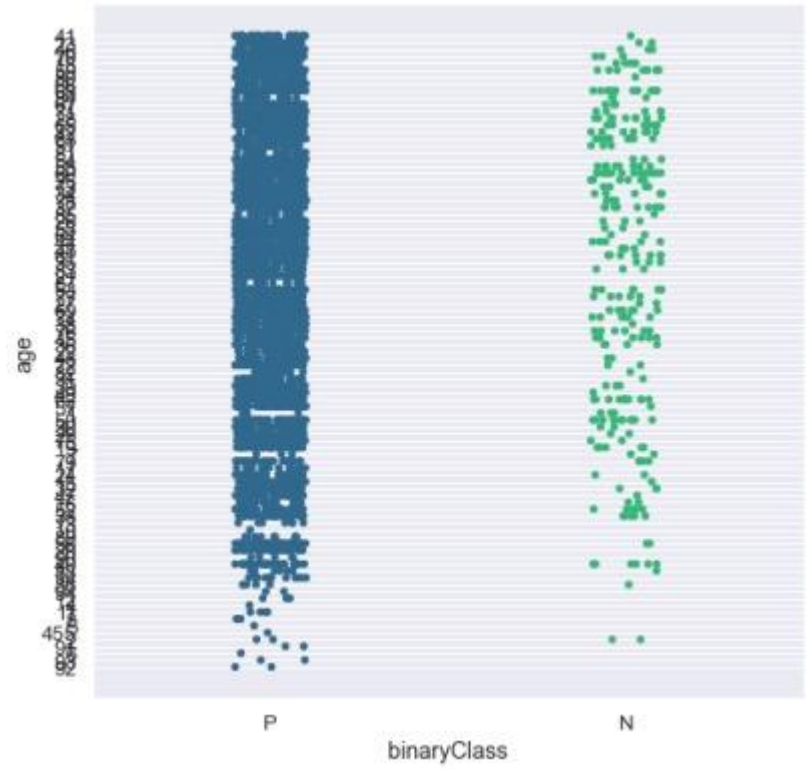
A joint plot is created for the variables 'TSH' (Thyroid-stimulating hormone) and 'binaryClass' using Seaborn's jointplot() function. This plot depicts the relationship between these two variables through a scatter plot, with marginal histograms for each variable's distribution.



Another joint plot is generated for the variables 'age' and 'TT4' (Total thyroxine) using Seaborn's jointplot() function. Similar to the previous joint plot, this visualization explores the correlation between age and TT4 levels.



Lastly, a strip plot is created for the variables 'binaryClass' and 'age' using Seaborn's stripplot() function. This plot illustrates the distribution of ages across different binary classes, providing insights into any potential age-related patterns or trends within each class.



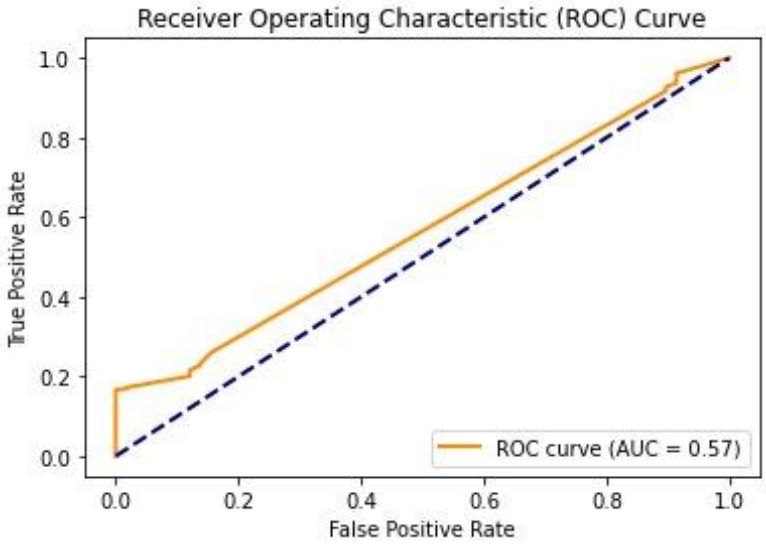
PART-3 The code begins by importing necessary libraries for evaluation, including sklearn.metrics for performance metrics and matplotlib.pyplot for visualization. It then computes and displays a confusion matrix, offering insights into the model's classification performance across different classes. Additionally, a heatmap is generated to visually represent the confusion matrix.

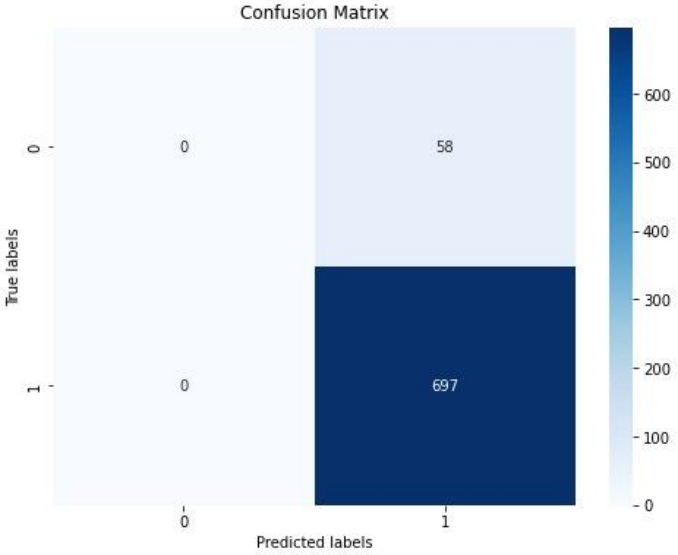
Following the confusion matrix, a classification report is printed, offering a detailed breakdown of precision, recall, F1-score, and support for each class, further illuminating the model's classification performance.

Subsequently, the logistic regression model is trained using the LogisticRegression class from scikit-learn. Predicted probabilities are computed for the test data, and the target variable is converted to a binary format for further analysis.

The code proceeds to compute the Receiver Operating Characteristic (ROC) curve and calculate the Area Under the Curve (AUC) score, providing a comprehensive evaluation of the model's discriminatory ability.

In addition to evaluation metrics, the code conducts feature importance analysis by extracting and sorting the absolute values of logistic regression coefficients. The top features influencing the target variable (e.g., hypothyroidism or hyperthyroidism) are identified and printed, offering insights into the factors driving the model's predictions.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.57)
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix
True labels
Predicted labels
0
58
0
697
0
1
600
500
400
300
200
100
0

```
AUC Score: 0.5669618562311384
Top 5 Features Influencing Hypothyroidism or Hyperthyroidism:
              Feature  Coefficient
0        on 1hyroxine     1.395506
7   query hypo1hyroid     1.146811
4            pregnan1     0.961126
10             goi1re     0.848355
5     1hyroid surgery     0.607810
```

```
[[  0  58]
 [  0 697]]
```

```
Classification Report:
              precision    recall  f1-score   support

           N       0.00      0.00      0.00        58
           P       0.92      1.00      0.96       697

    accuracy                           0.92       755
   macro avg       0.46      0.50      0.48       755
weighted avg       0.85      0.92      0.89       755
```

PART-4 Streamlit web application aimed at facilitating the prediction and analysis of thyroid disorders. Leveraging various data visualization techniques and user interaction components, the application offers a comprehensive exploration of a dataset comprising patient health metrics. Below is a detailed breakdown of the features and functionalities encapsulated within the application:

1. Data Loading and Preprocessing:

- The application begins by loading the dataset from a CSV file named "predictive_final_dataset.csv" and performing necessary preprocessing steps.

- It replaces missing values denoted as '?' with NaN (Not a Number) for consistency in data handling.

2. User Interface Design:

- Streamlit's capabilities are employed to design an intuitive user interface.

- The page title, icon, and layout are configured for branding and aesthetics.

- The interface includes descriptive headers, subheaders, and markdown text to provide context and guidance to users.

3. User Input Collection:

- Users can input their health metrics through interactive widgets such as sliders and checkboxes conveniently placed in the sidebar.

- Sliders allow users to input their age within a specified range.

- Radio buttons enable users to select their gender (Male/Female).

- Checkboxes are provided for binary variables related to various health conditions and treatments, allowing users to indicate their status (Yes/No).

4. Data Visualization:

- Various visualization techniques are employed to represent different aspects of the dataset:

- *Bar Charts:* Used to visualize the distribution of people having thyroid by sex.

- *Box Plots:* Illustrate the distribution of numerical features by thyroxine medication status.

- *Pie Charts:* Depict the percentage distribution of different categories within specific feature variables.

- *Correlation Analysis:* Displays the correlation of numerical features with the target variable (on thyroxine).

5. Insights and Analysis:

- Descriptive text accompanies each visualization to provide insights and analysis.

- Correlation analysis helps identify relationships between health metrics and the likelihood of being on thyroxine medication.

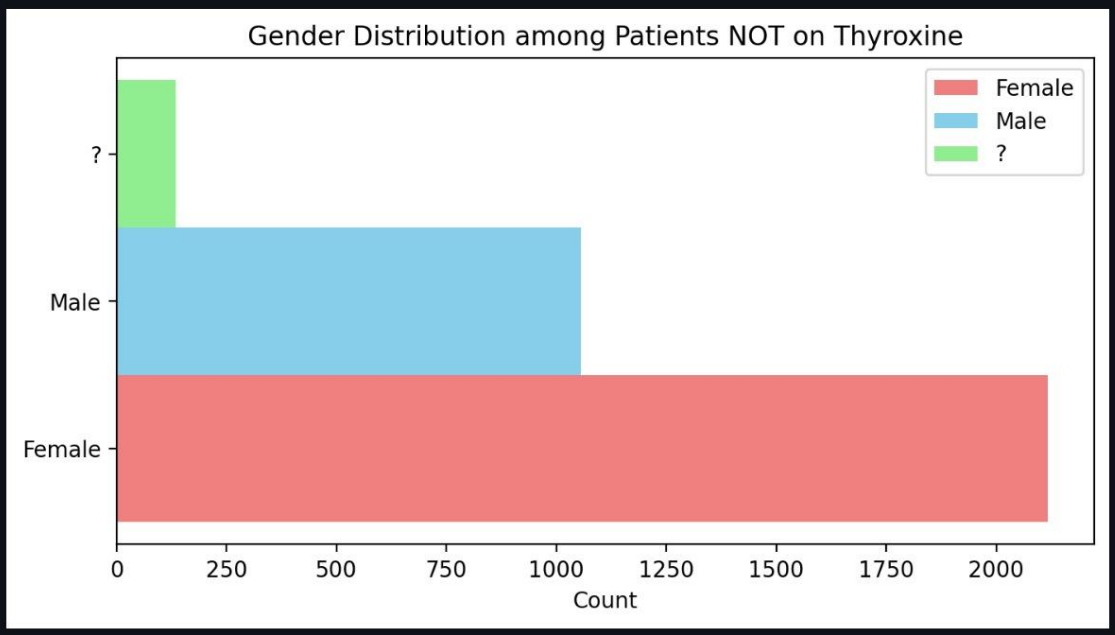
- Distribution comparisons based on medication status aid in understanding the impact of thyroxine treatment on various health parameters.

- Box plots offer detailed insights into the distribution and variability of numerical features among individuals based on their thyroxine medication status.

The application aims to serve as a valuable tool for healthcare professionals and researchers in the field of thyroid disorders.

- By facilitating data exploration, visualization, and analysis, it enables users to gain insights into patient demographics, treatment status, and correlations among health metrics.

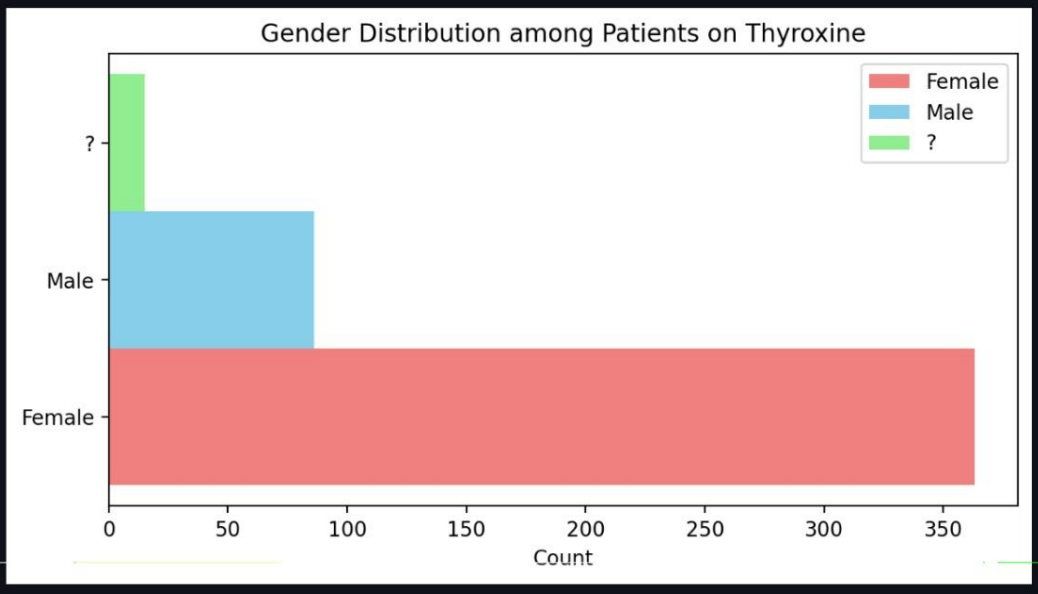
- The intuitive user interface and interactive features enhance usability, making it accessible to a wide range of users for research, analysis, and decision-making purposes.



Male: 1056

Female: 2117

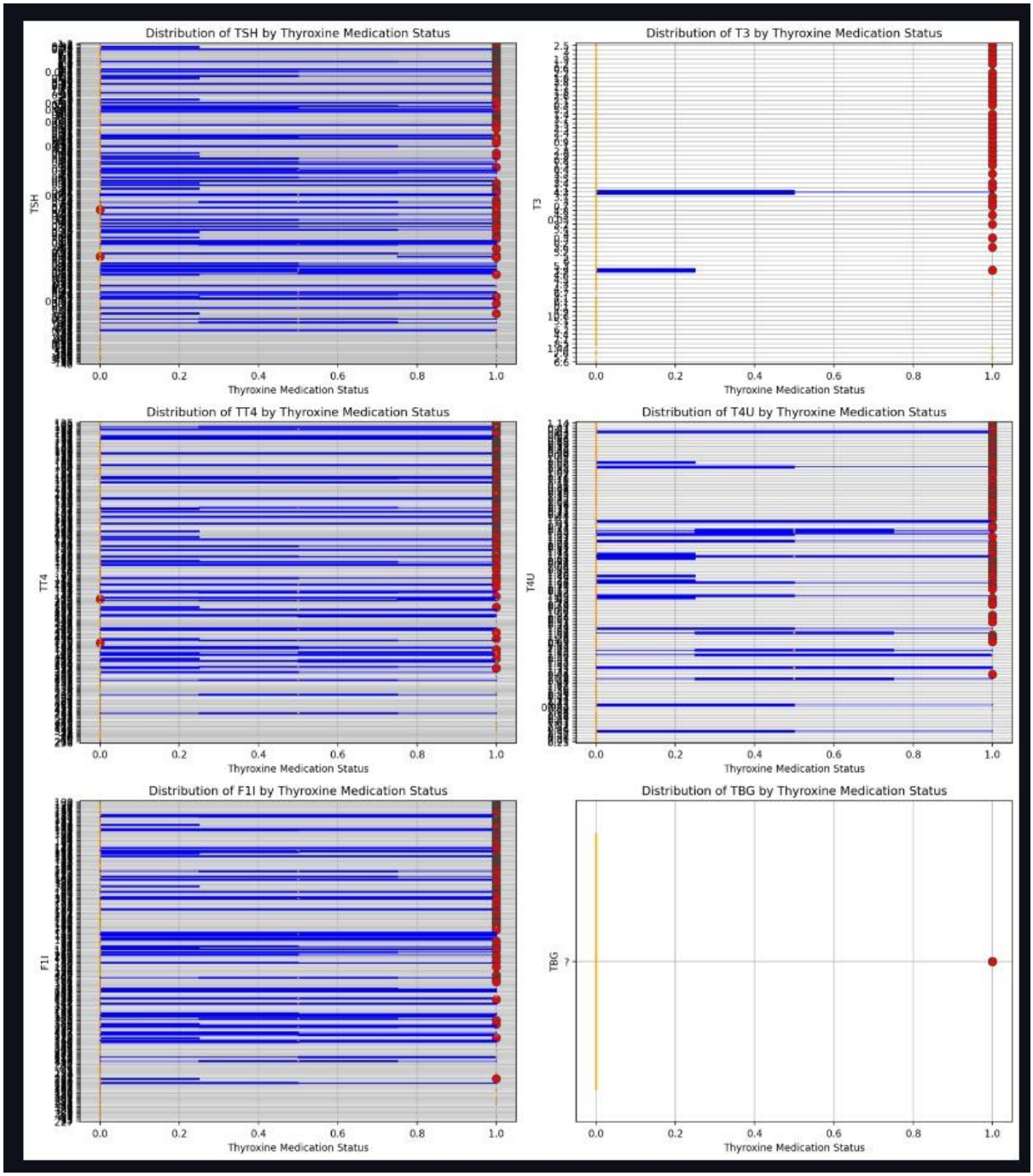
?: 135

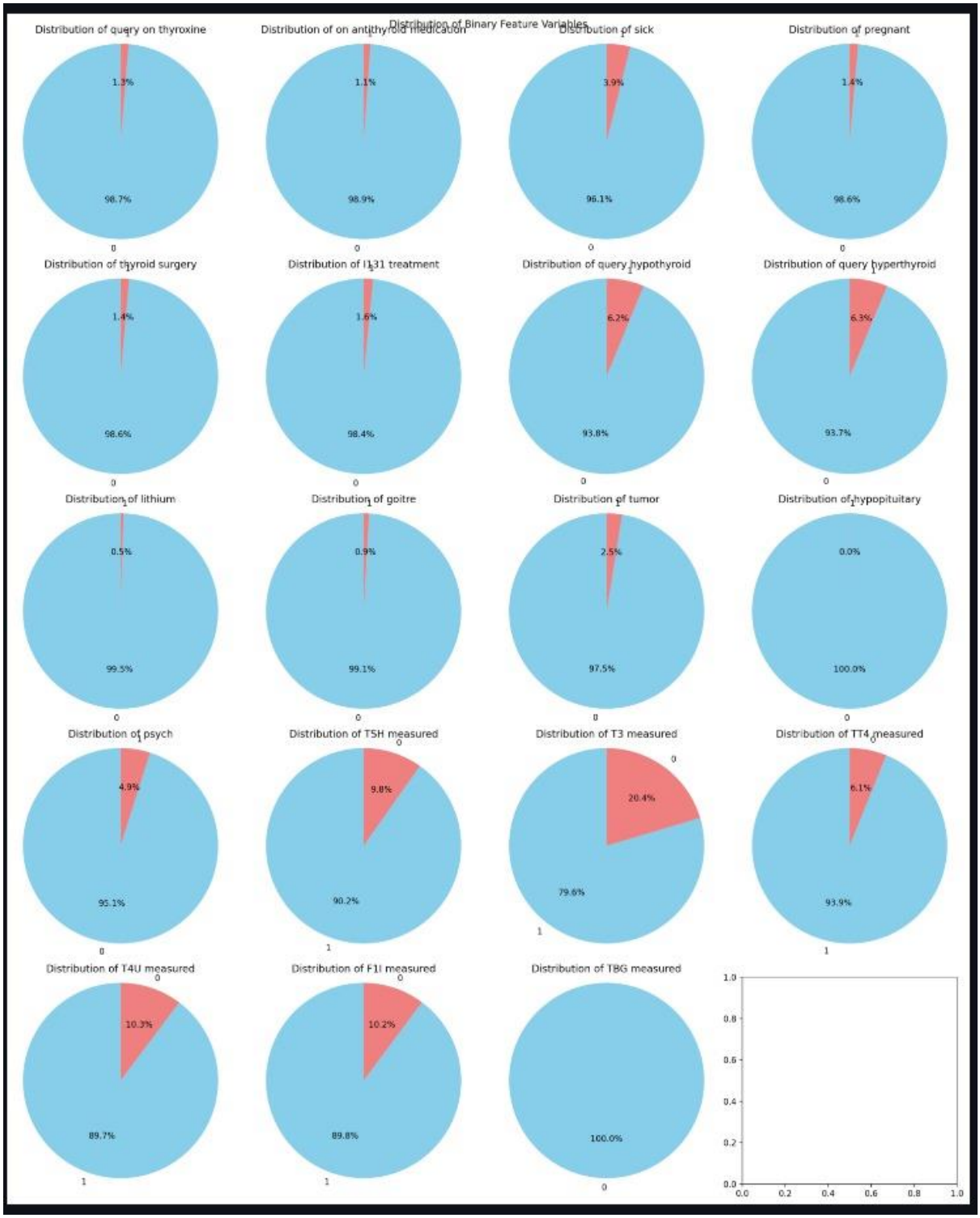


Male: 86

Female: 363

?: 15


Distribution of TSH by Thyroxine Medication Status
TSH
Thyroxine Medication Status
Distribution of T3 by Thyroxine Medication Status
T3
Thyroxine Medication Status
Distribution of TT4 by Thyroxine Medication Status
TT4
Thyroxine Medication Status
Distribution of T4U by Thyroxine Medication Status
T4U
Thyroxine Medication Status
Distribution of F1I by Thyroxine Medication Status
F1I
Thyroxine Medication Status
Distribution of TBG by Thyroxine Medication Status
TBG
Thyroxine Medication Status
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of Binary Feature Variables
Distribution of query on thyroxine
1.3%
98.7%
Distribution of on antithyroid medication
1.1%
98.9%
Distribution of sick
3.9%
96.1%
Distribution of pregnant
1.4%
98.6%
Distribution of thyroid surgery
1.4%
98.6%
Distribution of I131 treatment
1.6%
98.4%
Distribution of query hypothyroid
6.2%
93.8%
Distribution of query hyperthyroid
6.3%
93.7%
Distribution of lithium
0.5%
99.5%
Distribution of goitre
0.9%
99.1%
Distribution of tumor
2.5%
97.5%
Distribution of hypopituitary
0.0%
100.0%
Distribution of psych
4.9%
95.1%
Distribution of TSH measured
9.8%
90.2%
Distribution of T3 measured
20.4%
79.6%
Distribution of TT4 measured
6.1%
93.9%
Distribution of T4U measured
10.3%
89.7%
Distribution of FTI measured
10.2%
89.8%
Distribution of TBG measured
100.0%