Thyroid Disease Classification

Kaggle:

https://www.kaggle.com/rishabh458

Github:

https://github.com/Rishabh-a-git/D606-Thyroid_Classification.git

Team A:
Gaurav A Singh
Vaishnavi V Mane
Rishabh Anand

Thyroid Disease:

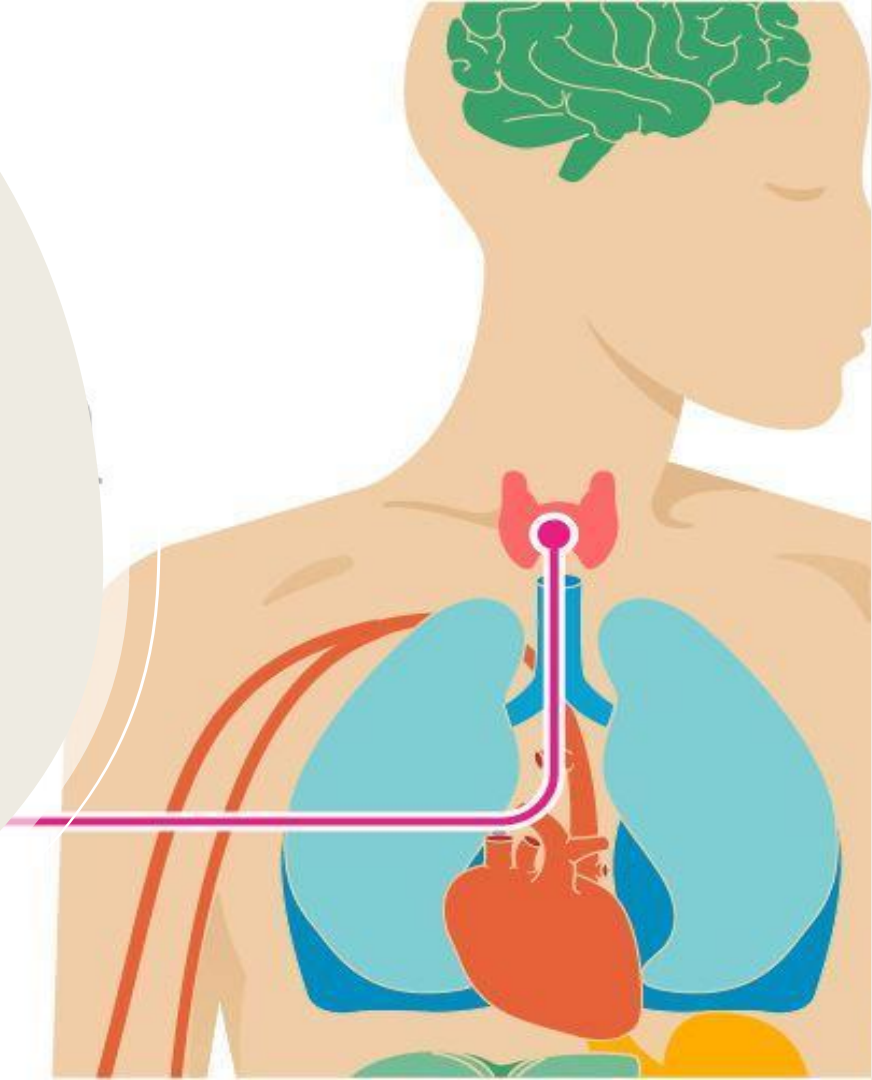
- Range of disorders affecting the thyroid gland.
- A crucial organ for regulating metabolism and maintaining overall health.
- Symptoms includes fatigue, weight changes, and mood disturbances.

Importance of Early Diagnosis:

- Subtle symptoms that can be easily overlooked.
- It can affect your metabolism, mental functions, energy level and bowel movements.

Objective :

- To develop a machine learning model capable of classifying patient into different thyroid conditions.
- This model will serve as a valuable tool for healthcare professionals in making accurate and timely diagnosis.



Literature Review

- **Thyroid Hormone Levels:** TSH (thyroid-stimulating hormone), T3 (triiodothyronine), TT4 (total thyroxine), and T4U (thyroxine-binding globulin) directly assess thyroid hormone levels. [1]

- **Symptoms and Conditions:** Goiter (enlarged thyroid gland), lithium use, psychological conditions, and hypopituitarism (underactive pituitary gland) are all signs or symptoms that can indicate thyroid dysfunction. [1]

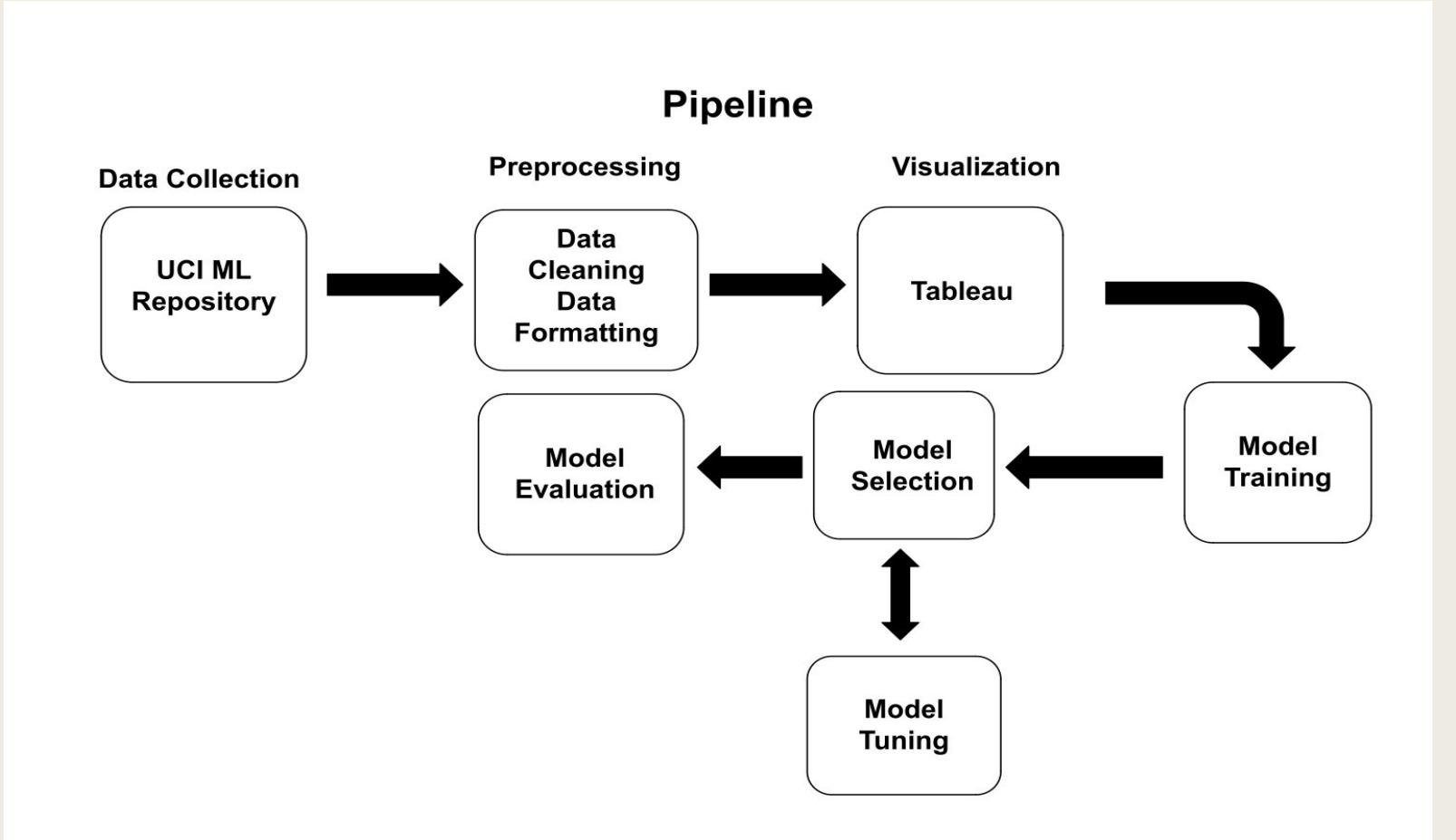
- **Other Factors:** Pregnancy can alter thyroid hormone levels, and age can influence thyroid function [3].

Literature Review

Study No.	Authors	Year	Classification- Class	Algorithms	Accuracy
1.	Shengjun Ji	2024	Hypothyroid, No condition, Increased binding protein, Compensated hypothyroid, Concurrent non-thyroidal illness	RF, GBM, AdaBoost etc.	RF-97% GBM-97% Adaboost-58%
2.	Rituraj Dixit; Madhuri A. Tayal; SarabjeetSingh Bedi; Shailesh Saxena	2023	4 classes Classification	Feature engineering with PCA, Decision Tree	TensorFlow-92.36% Decision Tree-87.67 %
3.	Khalid salman and Emrullah Sonuç	2021	Hyperthyroid, Hypothyroid, No condition,	Decision Tree, Random Forest etc.	DT- 98.4 % RF- 98.93 %
4.	Amulya.R. Rao; B.S. Renuka	2020	Stage (Major, Minor Critical) or No stage	Decision Tree and Naive Bayes	Various levels of precision and accuracy.

[5]

We developed a multi-class (seven) classification model of thyroid disease


Pipeline
Data Collection
Preprocessing
Visualization
UCI ML Repository
Data Cleaning Data Formatting
Tableau
Model Training
Model Selection
Model Evaluation
Model Tuning

Thyroid Dataset

The dataset comprises clinical data related to thyroid disease, encompassing various features such as patient demographics, medical history, and lab test results.

Dataset:

- Consists over 9000 instances
- Consists of 31 features

The ‘target’ column in the dataset represents various thyroid conditions.

Source:
https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data

Dataset

- **Missing Value Handling:**
  - Removed irrelevant columns.
  - Filtered outliers (age > 100).
  - Imputed missing values in "sex" with the most frequent value.
  - Imputed missing test hormone values (T3, T4, TSH) using group means based on age.
  - Filled remaining missing values in T4U and FTI with column means.

- **Feature Engineering:**
  - Created a new target variable "class" with descriptive labels for easier interpretation. Leaning on a research paper's findings, the code creates a new "class" column that maps target codes into more descriptive labels.
  - Added a "Patient_ID" for potential tracking.
  - Encoded categorical features using label encoding and pandas.get_dummies.

- **Data Preparation:**
  - Dropped unnecessary columns after creating new features.

Cleaned Data Snippet

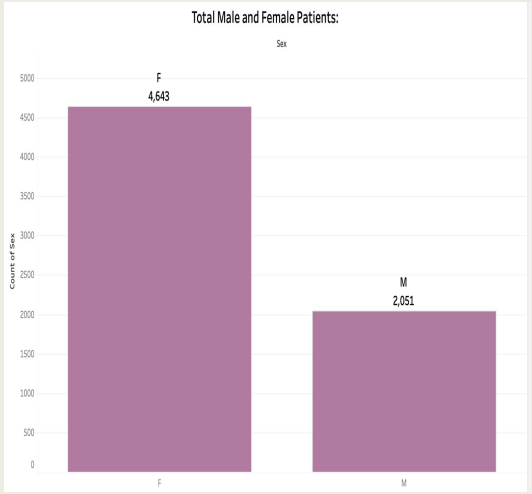
	age	TSH_x	T3_x	TT4_x	T4U	FTI	class	Patient_ID	sex_M	on_thyroxine_t	...	lithium_t	goitre_t	tumor_t	hypopituitary_t	psych_t
0	32	4.948343	2.232518	115.182155	0.984705	113.952402	Miscellaneous	5	False	False	...	False	False	False	False	False
1	63	68.000000	1.853211	48.000000	1.020000	47.000000	Hypothyroid	19	False	True	...	False	False	False	False	False
2	36	1.500000	2.400000	90.000000	1.060000	85.000000	No Condition	20	False	False	...	False	False	False	False	False
3	40	1.200000	2.300000	104.000000	1.080000	96.000000	No Condition	22	False	False	...	False	False	False	False	False
4	40	5.900000	2.100000	88.000000	0.840000	105.000000	No Condition	23	False	False	...	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6689	64	0.810000	1.853211	31.000000	0.550000	56.000000	General Health	9150	True	False	...	False	False	False	False	False
6690	60	0.180000	1.853211	28.000000	0.870000	32.000000	General Health	9154	True	False	...	False	False	False	False	False
6691	64	6.925247	1.853211	44.000000	0.530000	83.000000	Binding Protein	9155	True	False	...	False	False	False	False	False
6692	36	4.948343	2.232518	84.000000	1.260000	67.000000	Binding Protein	9159	False	False	...	False	False	False	False	False
6693	69	3.946271	1.752174	113.000000	1.270000	89.000000	Binding Protein	9166	True	False	...	False	False	False	False	False

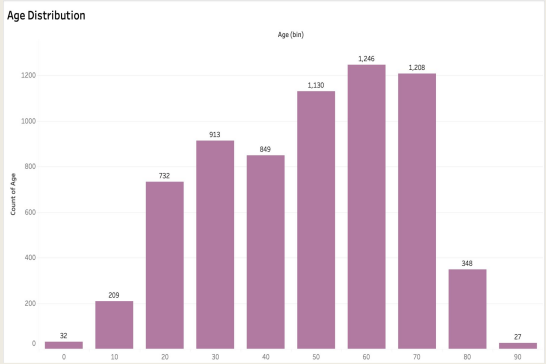
6694 rows × 28 columns

Data Visualization

Total Rows after Data Cleaning: 6694

Thyroid Disease Classification into 7 different classes





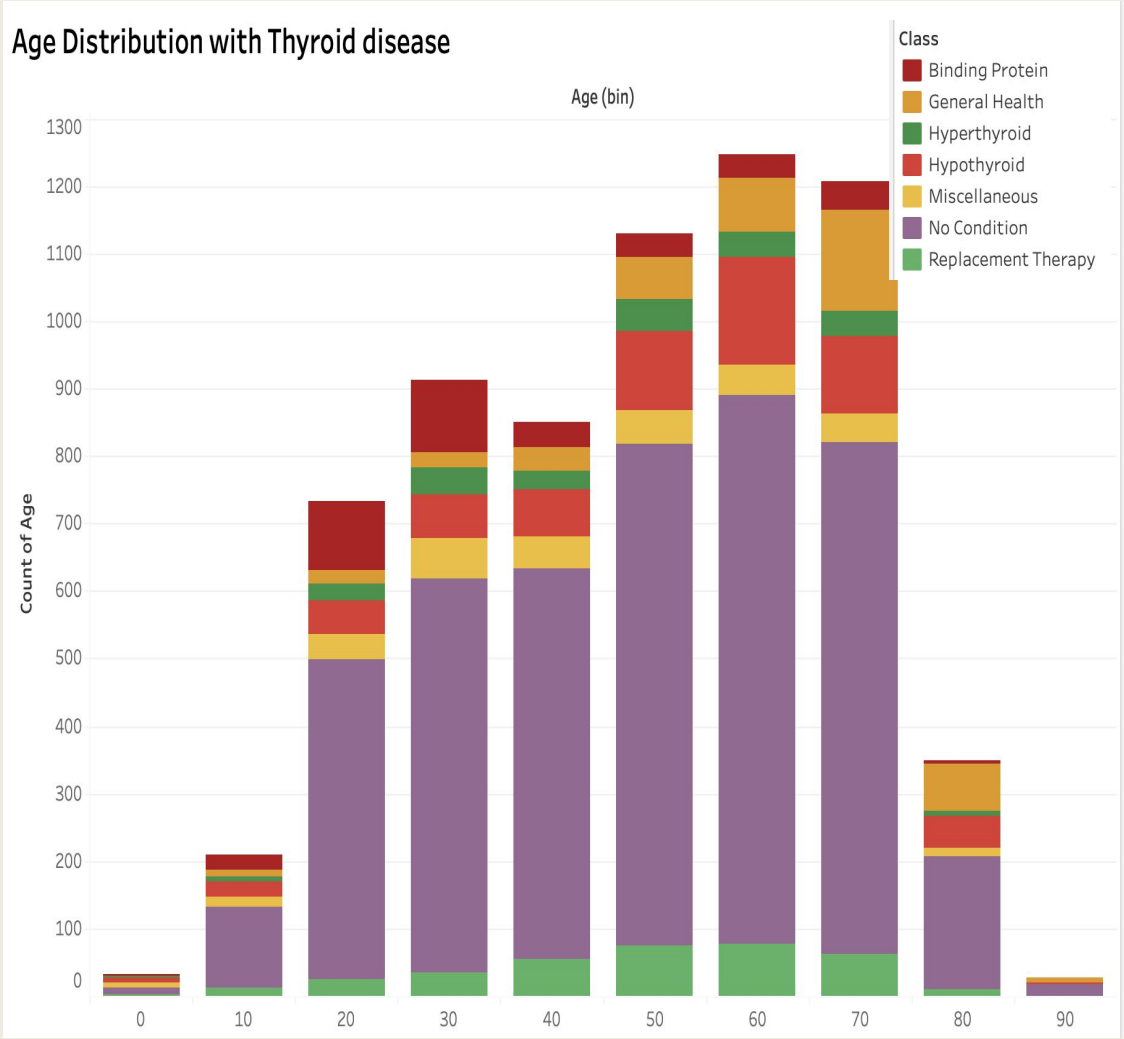
Age Distribution with Thyroid Disease:

Dataset has maximum patient between the age group of 20–70.

Hypothyroid cases increases with age and are dominant between age group of 50–70.

Hyperthyroid does not follow any specific trend across age

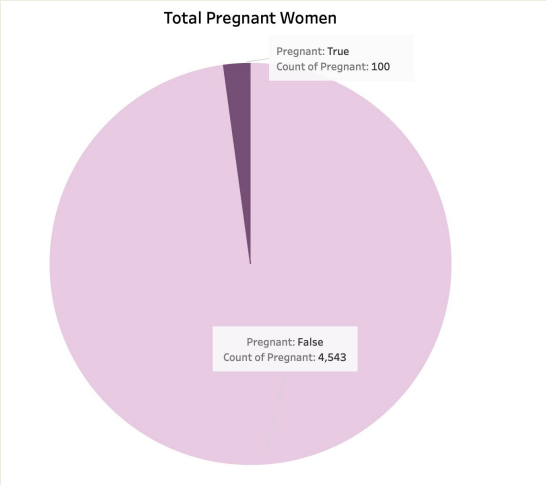
Binding Protein can be seen in younger age group between 20–30.



Studying Data for Thyroid Condition in Pregnant Women

Research States: Thyroid dysfunction in pregnant women including hypothyroidism and hyperthyroidism requires close monitoring and treatment as warranted.

There are 100 Pregnant females out of total 4643 females.



Almost 60% suffer from Binding Protein Condition.

Pregnant Women Classification

Class	
Binding Protein	69
Hyperthyroid	3
Miscellaneous	3
No Condition	23
Replacement Therapy	2

Test, train and validation

- It splits the pre-processed data (X – features, y – target class labels) into:

1. Training (X_train, y_train). (80%)
2. Test set (X_test, y_test). (20%)

- After applying Smoteenn training dataset was further split into :

1. Training (X_train, y_train) (75%)
2. Validation set(X_val, y_val) (25%)

- Validation set helps evaluate model performance before final testing on the unseen test.

SMOTEENN

SMOTEENN is used from the imblearn library to address potential class imbalance in the data.

SMOTEENN combines oversampling (SMOTE) with under-sampling (ENN) techniques to create a more balanced dataset.

Considerable care was taken to apply smoting only to training and validation dataset.

```
# Splitting the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#Shapes of datasets
print(f"Train set shape: {X_train.shape},Test set shape: {X_test.shape}")
```

```
Train set shape: (5355, 27),Test set shape: (1339, 27)
```

```
from imblearn.combine import SMOTEENN
# SMOTEENN to balance the classes
smoteenn = SMOTEENN(random_state=42)
X_train, y_train = smoteenn.fit_resample(X_train, y_train)
```

```
[ ] # Further split the train set into train and validation sets
X_train, X_val, y_train, y_val = train_test_split(X_train, y_train, test_size=0.25, random_state=42)
print(f"Train set shape: {X_train.shape}, Validation set shape: {X_val.shape}, Test set shape: {X_test.shape}")
```

```
Train set shape: (13255, 27), Validation set shape: (4419, 27), Test set shape: (1339, 27)
```

Binding Protein - Improving F2 Score

Thyroxine Binding Globulin

TBG column had over 8000 null values.

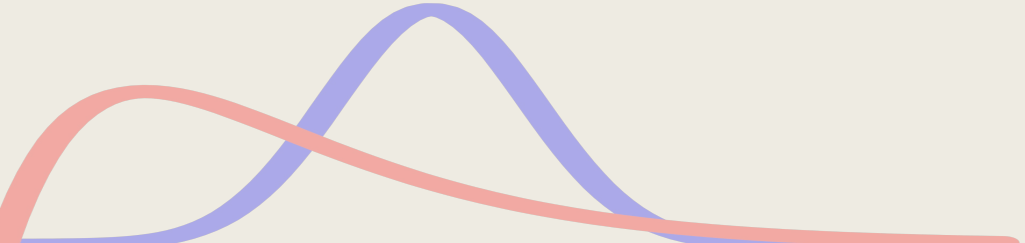
After model training: If TBG column Dropped–Binding Protein Classification were less. Less F2 score for binding protein

After imputing TBG null values by average values of the respected age and sex [6]---F2 score increased for binding protein

Age	Male (mg/dL)	Female (mg/dL)
1-5 days	2.2-5.9	2.2-5.9
1-11 months	3.1-5.6	3-5.6
1-9 years	2.5-5	2.5-5
10 to 19 years	2.1-4.6	2.1-4.6
Over age 20 years	1.2-2.5	1.4-3

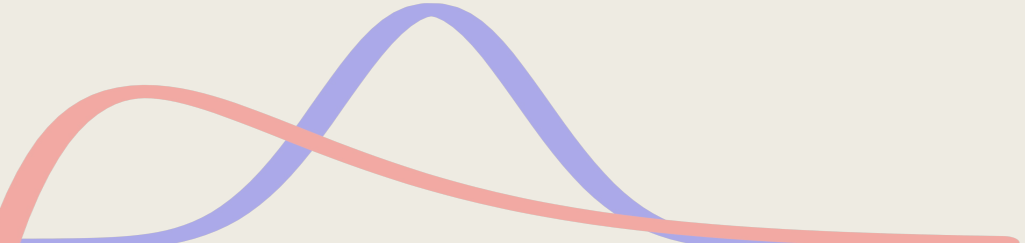
Candidate Models Evaluated using Cross-validation

- Logistic Regression
- Random Forest
- Gradient Boosting
- AdaBoost
- Decision Tree
- Gaussian Naive Bayes



Evaluation Metrics and Reasons behind it

- Why we chose Accuracy ?
- Why we chose Precision ?
- Why we chose Recall ?
- Why we chose F2 score over of F1 score ?

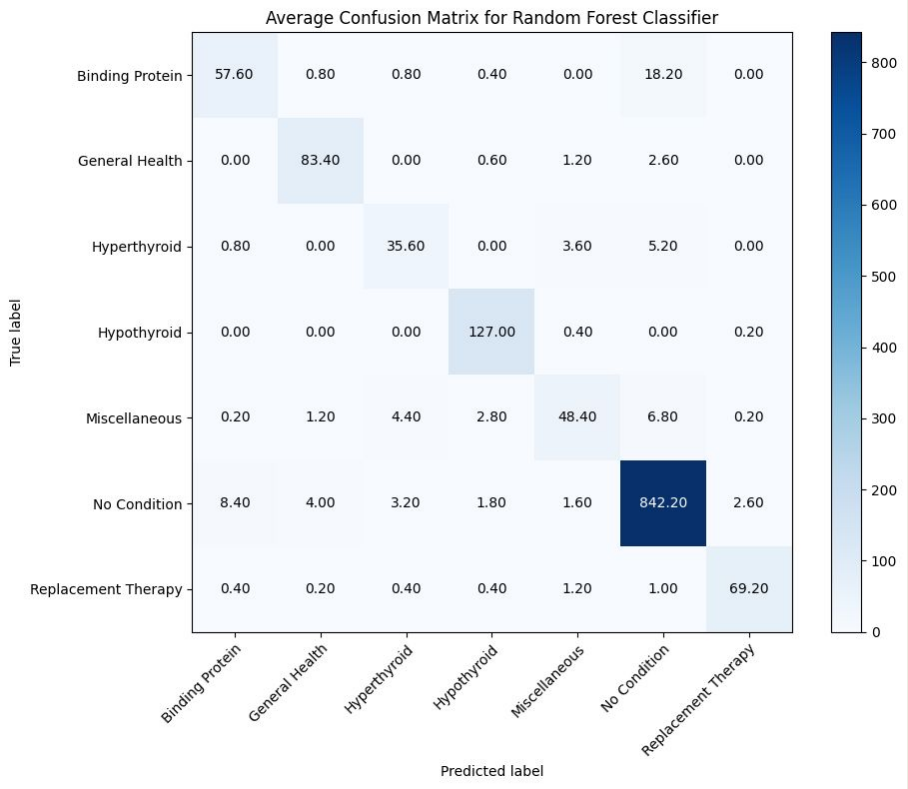


Results and Our Best Models

Model	Accuracy	F1 Score	Precision	Recall	F2 Score
Logistic Regression	0.623894	0.607185	0.598874	0.623894	0.616105
Random Forest	0.986062	0.985956	0.986017	0.986062	0.986001
Gradient Boosting	0.973451	0.973405	0.973431	0.973451	0.973424
AdaBoost	0.761504	0.762071	0.806381	0.761504	0.755664
Decision Tree	0.967035	0.966919	0.966875	0.967035	0.966980
Gaussian Naive Bayes	0.811062	0.808347	0.829273	0.811062	0.807292

Random Forest Classifier

Average Confusion Matrix and Other metrics:




```
Average F2 Score for Random Forest Classifier:
[0.76426208 0.94659121 0.78626726 0.98696558 0.77510648 0.97226265
 0.95219912]
```

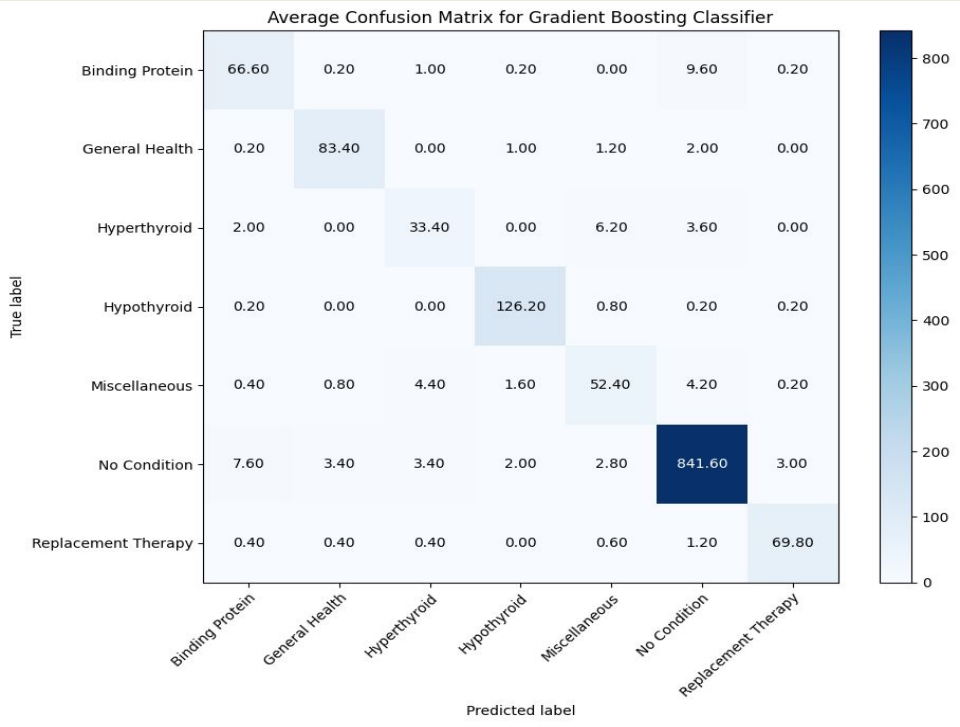
```
Classification Report:
                    precision    recall  f1-score   support

    Binding Protein       0.80      0.74      0.77        70
     General Health       0.94      0.93      0.93        98
      Hyperthyroid        0.75      0.82      0.79        40
      Hypothyroid         0.95      0.99      0.97       115
     Miscellaneous        0.88      0.75      0.81        71
      No Condition        0.96      0.97      0.97       871
Replacement Therapy       0.97      0.97      0.97        74

          accuracy                            0.94      1339
         macro avg        0.89      0.88      0.89      1339
      weighted avg        0.94      0.94      0.94      1339
```

Gradient Boosting Classifier

Average Confusion Matrix and Other metrics:




```
Average F2 Score for Gradient Boosting Classifier:
[0.85836013 0.95021901 0.7443499  0.98408194 0.81805509 0.97461735
 0.95728725]
```

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

    Binding Protein       0.87      0.86      0.86        70
     General Health       0.95      0.92      0.93        98
      Hyperthyroid        0.70      0.80      0.74        40
      Hypothyroid         0.93      1.00      0.96       115
     Miscellaneous        0.88      0.73      0.80        71
      No Condition        0.97      0.97      0.97       871
Replacement Therapy       0.96      0.97      0.97        74

         accuracy                             0.95      1339
        macro avg       0.89      0.89      0.89      1339
     weighted avg       0.95      0.95      0.95      1339
```

Decision Tree Classifier:

Average Confusion Matrix and Other Metrics:

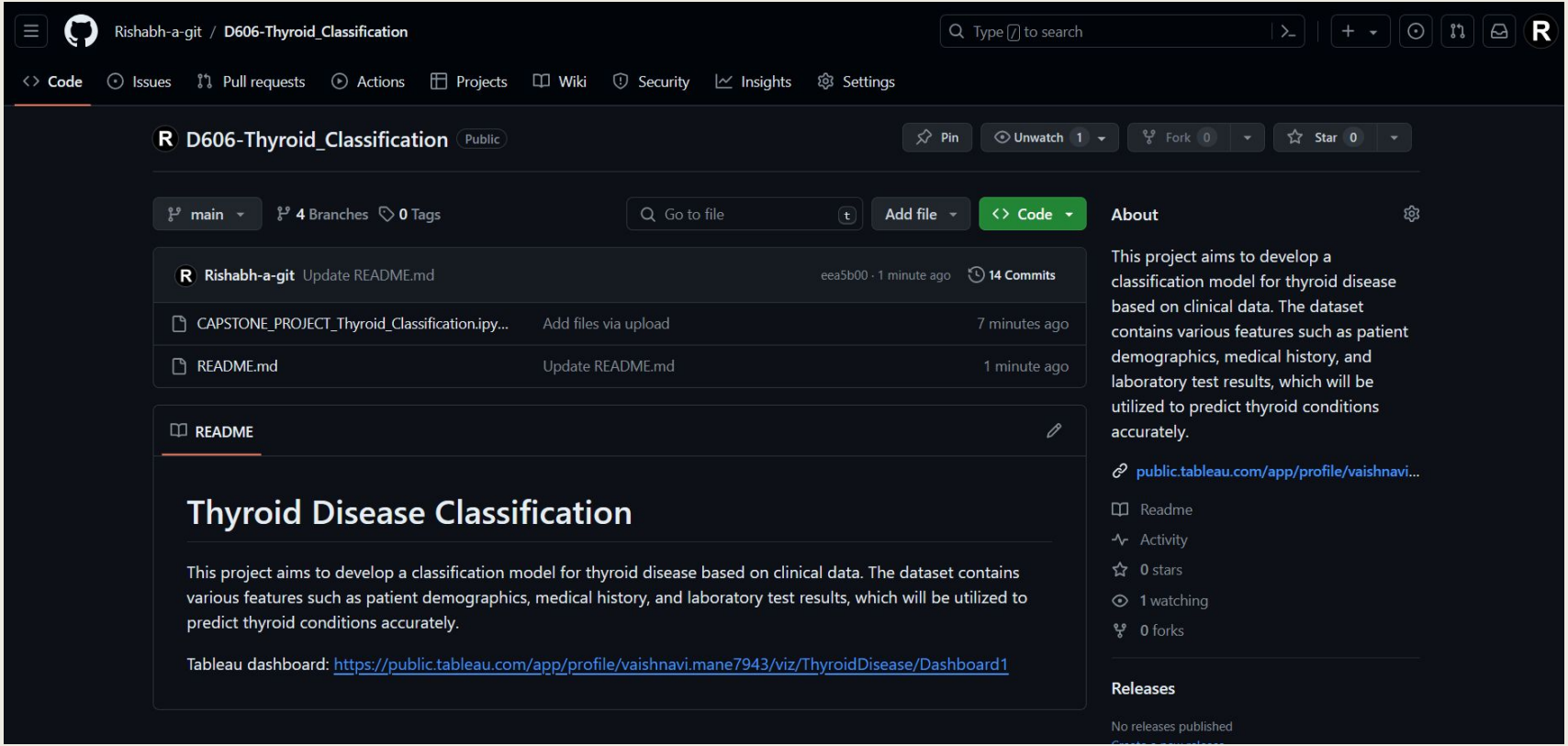

```
Average F2 Score for Decision Tree Classifier:
[0.80132888 0.9439728  0.71430291 0.96364657 0.77781339 0.9662366
 0.92375794]
```

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

    Binding Protein       0.70      0.73      0.71        70
     General Health       0.95      0.94      0.94        98
      Hyperthyroid        0.70      0.70      0.70        40
      Hypothyroid         0.93      0.95      0.94       115
     Miscellaneous        0.83      0.75      0.79        71
      No Condition        0.97      0.97      0.97       871
Replacement Therapy       0.90      0.95      0.92        74

          accuracy                            0.93      1339
         macro avg        0.85      0.85      0.85      1339
      weighted avg        0.93      0.93      0.93      1339
```

Github Snapshot



Future Scope

Data Collection and Quality Improvement: Enhance data collection protocols to ensure comprehensive and accurate recording of variables such as TBG (Thyroxine-Binding Globulin) to reduce missing values and improve data quality.

Deep Learning techniques or Ensemble methods: Models that combine multiple models, could lead to improved performance and robustness in thyroid disorder classification.

References

1. *A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques*. (2019, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9002284
2. Duntas, L. H., & Jonklaas, J. (2019). Levothyroxine dose adjustment to optimise therapy throughout a patient's lifetime. *Advances in Therapy*, *36*(S2), 30–46. https://doi.org/10.1007/s12325-019-01078-2
3. Leso, V., Vetrani, I., De Cicco, L., Cardelia, A., Fontana, L., Buonocore, G., & Iavicoli, I. (2020). The Impact of thyroid diseases on the working life of patients: A Systematic review. *International Journal of Environmental Research and Public Health*, *17*(12), 4295. https://doi.org/10.3390/ijerph17124295
4. Martinekuan. (n.d.). *Machine learning operations (MLOps) framework to upscale machine learning lifecycle with Azure Machine Learning - Azure Architecture Center*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/mlops-technical-paper
5. Salman, K., & Sonuç, E. (2021). Thyroid disease classification using machine learning algorithms. *Journal of Physics. Conference Series*, *1963*(1), 012140. https://doi.org/10.1088/1742-6596/1963/1/012140
6. Liess, B. D., MD. (n.d.). *Thyroid-Binding Globulin: reference range, interpretation, collection and panels*. https://emedicine.medscape.com/article/2089554-overview?form=fpf

Thank You