# Thyroid Class Classification

Ayomi Upeksha

2023-09-30

## 1.Inroduction

Millions of people throughout the world suffer from thyroid conditions, which frequently result in a variety of health problems. For successful treatment and better patient outcomes, thyroid problems must be identified early and correctly diagnosed. Machine learning presents a possible way to improve thyroid illness diagnosis in this era of cutting-edge technology. The goal of this investigation is to use machine learning to create a predictive model that will help medical professionals recognize thyroid diseases quickly and reliably.

The thyroid, a crucial component of the endocrine system, is in charge of regulating metabolism and safeguarding overall health. Thyroid dysfunction, including hyperthyroidism, hypothyroidism, thyroid nodules, and autoimmune illnesses, can have a substantial negative impact on a person's health. Unfortunately, these disorders' modest and occasionally overlapping symptoms make diagnosis challenging. So, determinations must therefore be based on sophisticated, complex data analysis.

The "MLDataR" package in R from the database has been used in this instance. It draws on data from the UCI Machine Learning repository and contains patient records.

## 2. Data Preprocessing

**Data Profiling :** There are 22 categorical variables, 6 numerical variables, and 3772 rows in this dataset. In order to better comprehend the data, the variable definition for this dataset is described below.

ThryroidClass: Patient's status (sick = 1 or negative=0)

patient_age: Age of the patient.

patient_gender: Female = 1 & Male = 0

presc_thyroxine : Whether thyroxine replacement prescribed 1=Thyroxine prescribed

queried_why_on_thyroxine : Indicate query has been actioned

presc_anthyroid_meds : Whether anti-thyroid medicine has been prescribed

sick : Sickness due to thyroxine depletion or over activity

pregnant : Whether the patient is pregnant

thyroid_surgery : Whether the patient has had thyroid surgery

radioactive_iodine_therapyI131 : Whether patient has had radioactive iodine treatment:

query_hypothyroid : Indicate under active thyroid query

query_hyperthyroid : Indicate over active thyroid query

lithium : Lithium carbonate administered to decrease the level of thyroid hormones

goitre : Indicate swelling of the thyroid gland

tumor : Indicate a tumor

hypopituitarism : Indicate a diagnosed under active thyroid

psych_condition : Whether a patient has a psychological condition

TSH_measured : A TSH level lower than normal indicates there is usually more than enough thyroid hormone in the body and may indicate hyperthyroidism

TSH_reading : Reading result of the TSH blood test

T3_measured : Linked to TSH reading - when free triiodothyronine rise above normal this indicates hyperthyroidism

T3_reading : Reading result of the T3 blood test looking for above normal levels of free triiodothyronine

T4_measured : Free thyroxine, also known as T4, is used with T3 and TSH tests to diagnose hyperthyroidism

T4_reading : Reading result of th T4 test

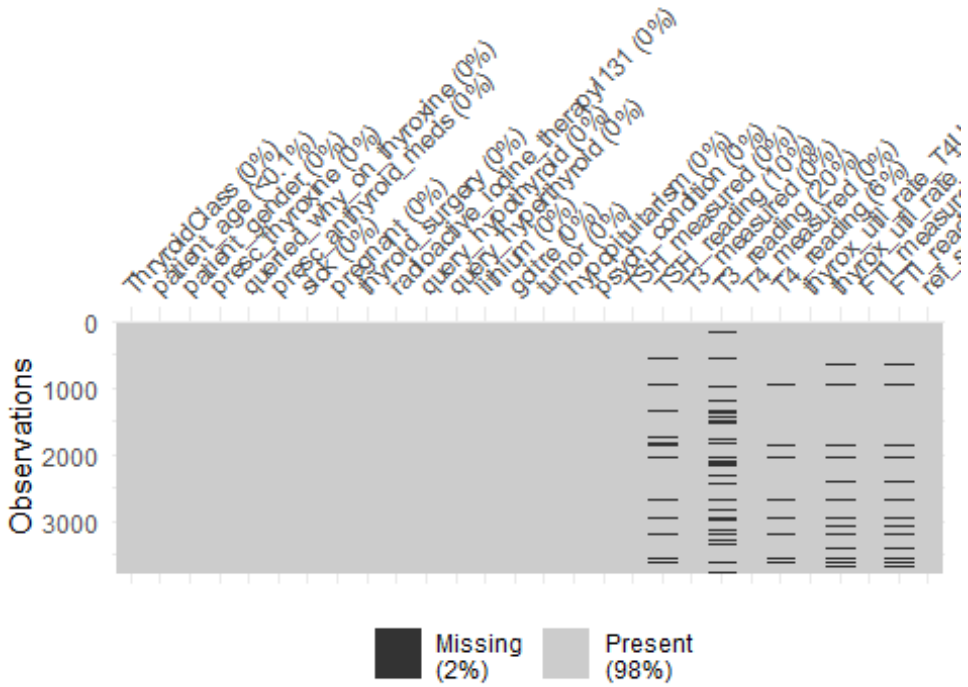
thyrox_util_rate_T4U_measured : Thyroxine utilisation rate

thyrox_util_rate_T4U_reading : Result of the test

FTI_measured : Measurement on the Free Thyroxine Index (FTI)

FTI_reading : Result of the test mentioned above

ref_src : Referral source of the patient

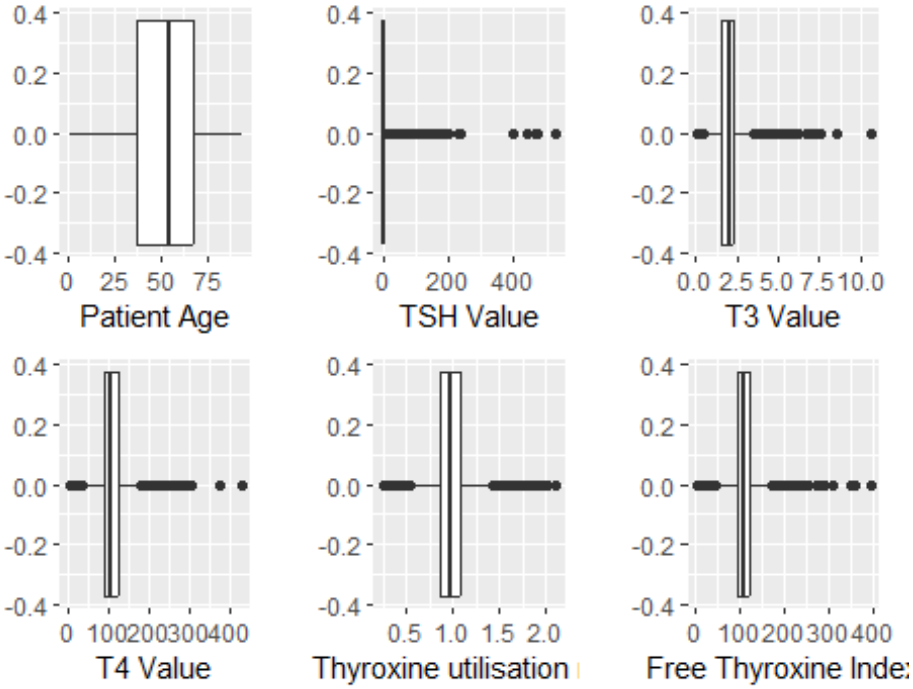



**Data Cleansing :** According to the above missing value visualization figure, It can be seen that, 2% of missing values exists in TSH_reading, T3_reading, T4_reading, thyrox_util_rate_T4U_reading, FTI_reading and patient_age. Looking carefully at the dataset, it is clear that TSH_reading occurred some missing values, if TSH_measured is zero. The same procedure is applied for other remaining for variables as well. There are no important information cannot be generated TSH_measured, T3_measured, T4_measured, thyrox_util_rate_T4U_measured and FTI_measure. Hence, Those records are eliminated from the dataset.

The dataset size is decreased to 2751 after the elimination process, and variables such as TSH_measured, T3_measured, T4_measured, thyrox_util_rate_T4U_measured, and FTI_measure also removed from the dataset because these variables no longer need the dataset.

**Data Smoothing:** According to figure 2.1, there are numerous noticeable outliers in the TSH, T3, and T4 levels as well as the Free Thyroxine index. Therefore, without doing a thorough study, such observable outliers cannot be directly eliminated from the dataset.According to the TSH level, an increase of more than 400 is abnormal. Records that exceed TSH level 400 Can be eliminated from the dataset.[1]

Figure 2.1 One Dimensional Outlier Detection for numerical variables

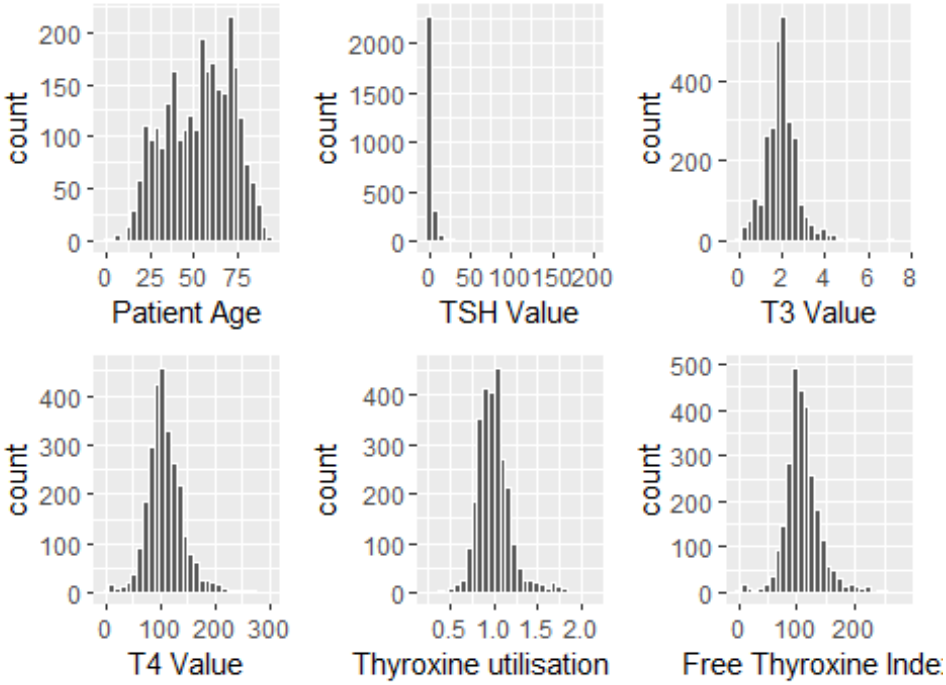



## 3. Exploratory Data Analysis

### 3.1 Distribution of Patient's age, TSH value, T3, T4, Thyroxine utilization rate, Thyroxine Index

In Figure 3.1, a prominent pattern emerges as we analyze the distributions. The majority of variables display a characteristic centered bell-shaped curve, indicative of a normal distribution. This suggests that data points tend to cluster around the mean, with few outliers. Notably, the distributions of TSH values and T3 values defy this trend. They exhibit positive skewness, signifying an elongated tail on the higher values side. This skewness could imply a higher incidence of health conditions associated with abnormal TSH and T3 levels. Consequently, these skewed distributions may require particular attention in subsequent analyses to identify potential outliers and health-related insights in this dataset.

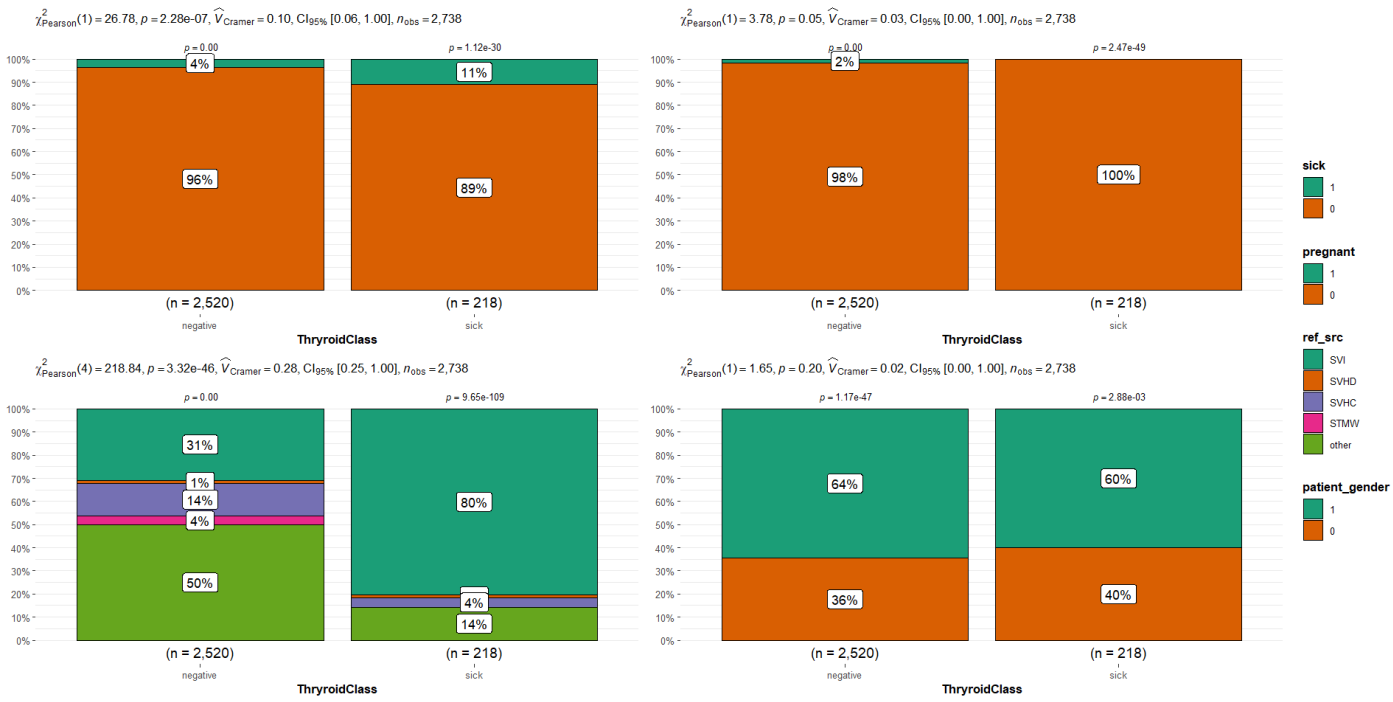
Figure 3.1 Distribution of numerical variables




### 3.2 Distribution of Patient's gender, sickness, pregnant status, Referral source and Hypopituitarism with Thyroid Class

In light of the multitude of categorical variables in this dataset, it becomes imperative to delve into their relationships with the categorical 'thyroid class' variable. To accomplish this, It has embarked on an investigative journey to uncover the intricate web of associations. The approach involved calculating pairwise Pearson correlation coefficients between these qualitative variables and the 'thyroid class,' which serves as the target variable. This endeavor not only aids in elucidating the interplay between categorical predictors and the thyroid class but also offers valuable insights into which variables might wield significant influence in predicting thyroid disease outcomes. This correlation analysis equips us with the means to discern which categorical attributes may exhibit stronger or weaker connections with the 'thyroid class', thus paving the way for more informed modeling and decision-making in the context of thyroid disease prediction and diagnosis.

Figure 3.2.1 Distribution of Thyroid class with Sikness, Referral source,pregnant & gender




In Figure 3.2, we discern meaningful correlations between categorical variables and the 'thyroid class.' Notably, 'sickness,' 'pregnant status,' and 'referral source' exhibit pronounced associations with thyroid class. Patients with sickness due to thyroxine depletion represent 11% of thyroid-positive cases, in contrast to 4% in the thyroid-negative group, emphasizing the impact of this condition on thyroid status. Intriguingly, 'pregnant status' does not seem to elevate thyroid disease risk significantly, as indicated by a comparable prevalence in both thyroid-positive and -negative cases. 'Referral source' is a pivotal factor, with 80% of thyroid-positive patients stemming from 'SVI,' compared to 31% from the thyroid-negative class. Notably, other referral categories, including 'SVDH,' 'SVHC,' 'STMW,' and 'other,' do not exert a discernible influence on thyroid-positive cases. Furthermore, 'Whether thyroxine replacement is prescribed' and 'active status under thyroid query' exhibit significant relationships with 'thyroid class,' underscoring their importance in predicting thyroid disease outcomes. These findings offer valuable insights into the multifaceted factors affecting thyroid status, guiding further exploration in thyroid disease analysis.

While gender status may not exhibit a particularly strong influence on thyroid sickness, an intriguing pattern emerges when we consider the proportions within each category. Notably, 40% of thyroid-positive cases occur, juxtaposed with 36% among negative cases. Similarly, the data for females and males reveal a mirror image of this distribution. This same analytical approach is systematically applied to all other variable pairs, with p-values is recorded to inform decision-making process. It is worth noting that these other variable pairs do not

exhibit correlations as strong as the previously mentioned variables. In fact, most other qualitative variables do not seem to significantly impact thyroid class status. Nonetheless, it remains prudent to include these variables in our model, as doing so may enhance model accuracy.

Figure 3.2.2 Distribution of Thyroid class with Hypopituitarism and Action status of Quirey

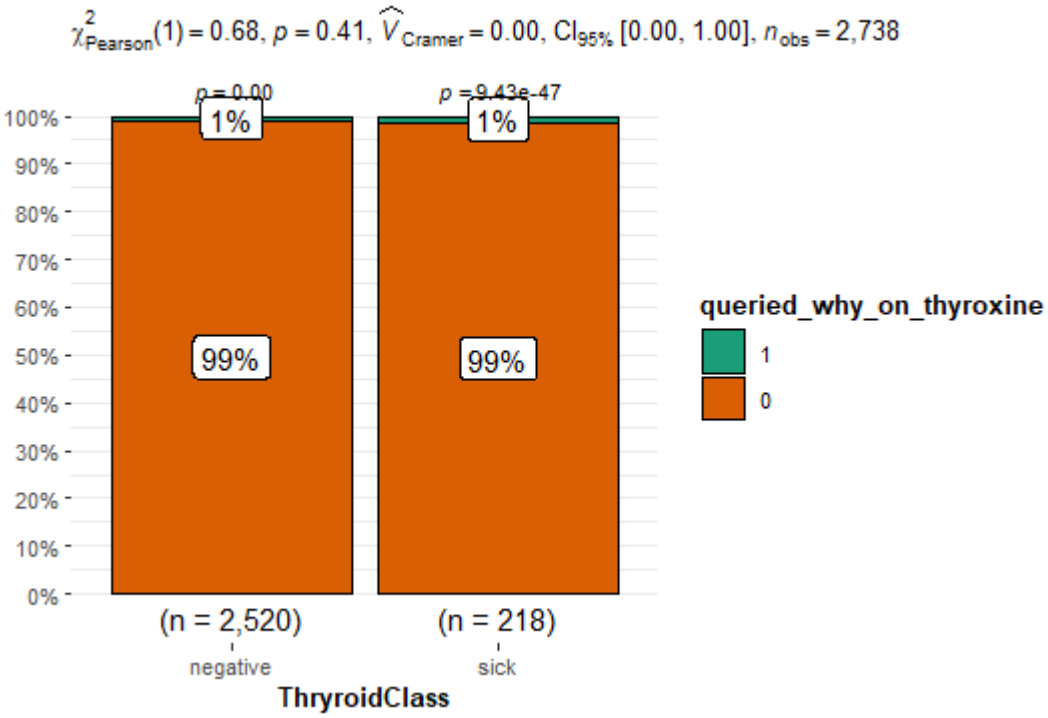



Figure 3.3 reveals that the presence or absence of query action does not exhibit any discernible relationship with thyroid class. Both thyroid-positive and negative categories demonstrate an almost identical proportion, with 99% in the 'actioned' status and 1% in the 'not actioned' status. Given this lack of distinction, it is advisable to exclude this variable from subsequent stages of the analysis.

## 4. Model Building & Evaluation

**Splitting Data:** After preprocessing, the dataset was reduced to 2738 samples and retained 22 columns. To facilitate model building, it was essential to split the dataset into training and testing sets, with an 80% allocation to the training set. Consequently, the training set comprises 2190 records, while the testing set consists of 548 records.

```
df_14 <- df_13 %>% mutate(id = row_number())

set.seed(123)

train_thy <- df_14 %>% sample_frac(.80)

test_thy <- anti_join(df_14, train_thy, by = 'id')
```

**Making Recipe:** In the concluding data processing step, we prioritize the normalization of quantitative variables linked to the thyroid class. Figure 3.1 highlights the need for this step, as individual distributions, particularly TSH values and patients' ages, exhibit disparities in their ranges. Normalization is essential to ensure that these variables are on a consistent scale, enabling more effective modeling and analysis.

```
thyroid_recipe <-

recipe(ThryroidClass ~ ., data = train_thy) %>% # Setting the formula

step_normalize(all_numeric_predictors()) %>% # Normalize every numeric predictor

step_zv(all_predictors()) %>% # Remove predictors with only one value (they are useless)

prep() # Apply the recipe
```

```
# Applying the recipe to training set

thyroid_training <- juice(thyroid_recipe) #extract the training dataset
```

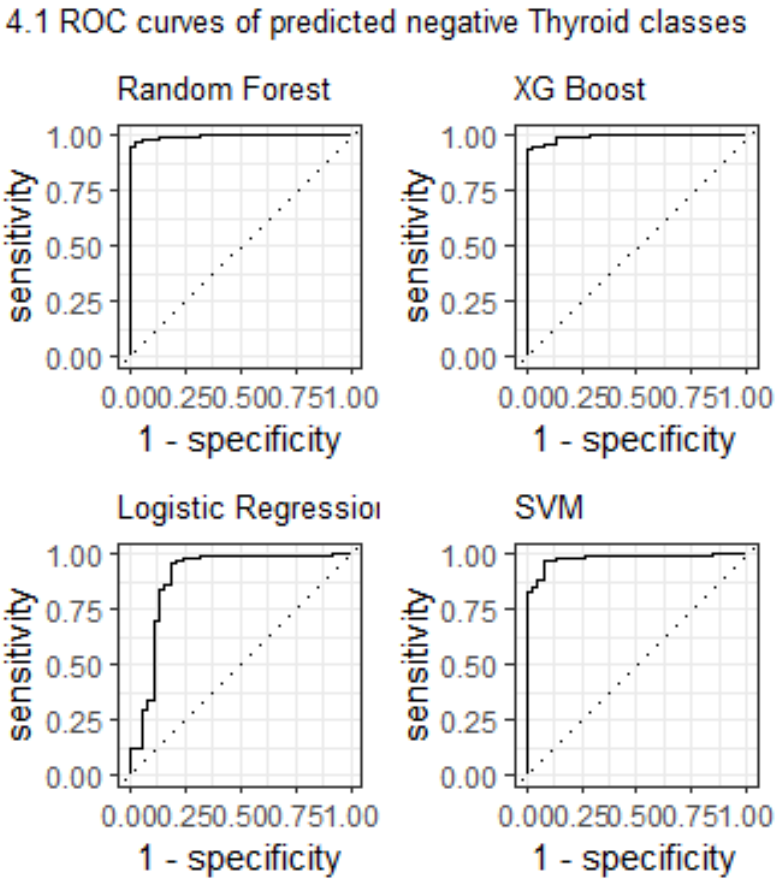
```
# perform the same task on the testing set

thyroid_testing <- thyroid_recipe %>% bake(test_thy) # bake uses for apply the same reci
pe for test data
```

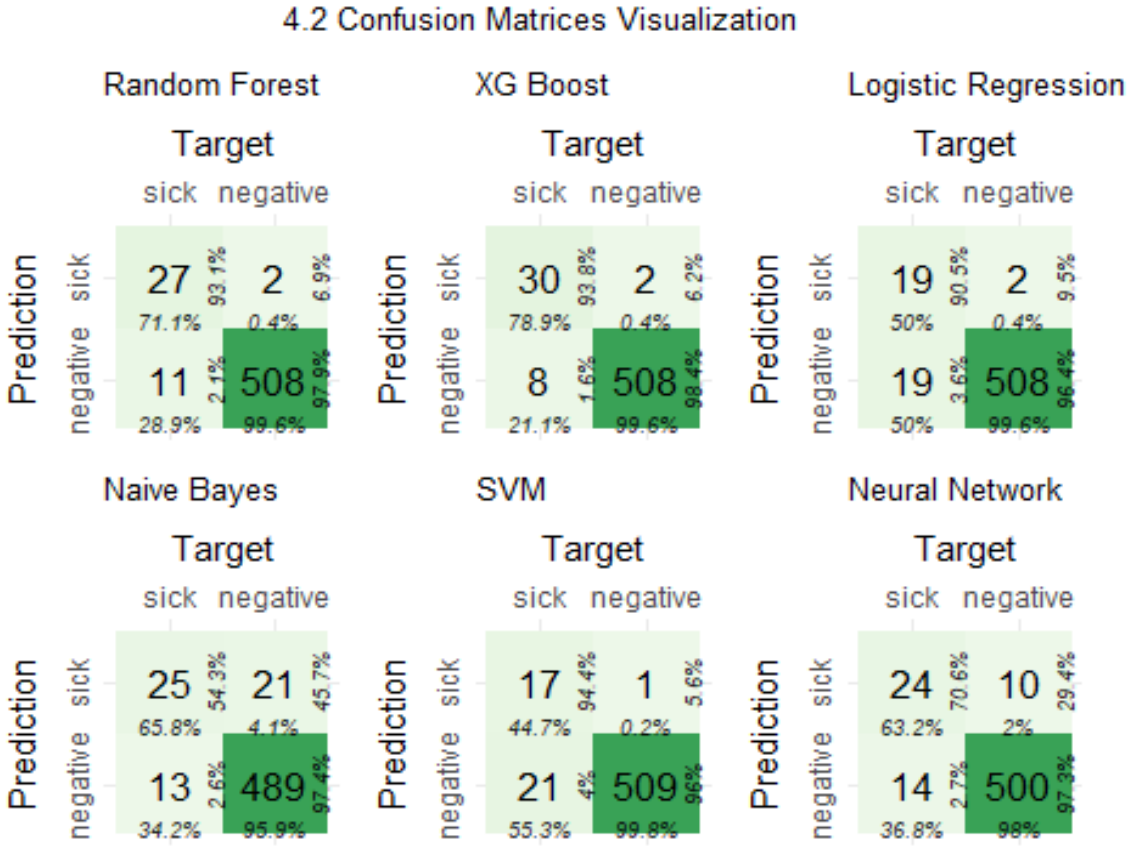
**Model Building:** Multiple models have been constructed and rigorously evaluated for their predictive accuracy, with a primary focus on discussing their distinctive features. The models developed for thyroid prediction encompass:

- Random Forest Model
- Extreme Gradient Boosting (XG Boost)
- Logistic Regression
- Naive Bayes Classifier

- Support Vector Machine
- FeedFoward Neural Network (FNN)




In Figure 4.1, we observe the ROC curves of the machine learning models developed for thyroid class predictions. Notably, all these curves consistently surpass the baseline centered line at 0.5. This signifies that the area under the curve (AUC) for each model exceeds 0.5, indicating their efficacy in distinguishing between thyroid-positive and thyroid-negative cases. It's worth noting that among the models evaluated, the Logistic Regression model exhibits a relatively lower AUC compared to the other curves, suggesting that while it still demonstrates predictive capability, alternative models may offer superior performance in this context.




This confusion matrices are presented in figure 4.2, shows the matrix representation of the prediction summary of the thyroid classes. The confusion matrix results of our thyroid classification model showcase its effectiveness in distinguishing between “Negative” and “Sick” thyroid cases.

**Random Forest Classifier:** In a more detailed examination of the confusion matrix, It has observed that the model correctly identified 508 cases as “Negative” (true negatives) and 27 cases as “Sick” (true positives). However, there were 11 instances where the model erroneously predicted “Sick” when the actual class was “Negative” (false positives). Additionally, the model misclassified 2 cases as “Negative” when they were actually “Sick” (false negatives). It correctly identified 71.10% of ‘Sick’ cases (sensitivity) and 99.6% of ‘Negative’ cases (specificity). The positive predictive value (precision) for ‘Sick’ cases is 93.10%, with a negative predictive value of 97.4%.

**XG Boost Classifier:** The model correctly identified 508 cases as 'Negative' (true negatives) and 30 cases as 'Sick' (true positives). However, there were 8 instances where the model incorrectly predicted 'Sick' when the actual class was 'Negative' (false positives), and 2 instances where it predicted 'Negative' when the actual class was 'Sick' (false negatives). This performance demonstrates the model's strength in correctly identifying 'Sick' cases, but there is room for improvement in reducing false positives.

**Logistic Regression Classifier:** The model correctly identified 508 cases as 'Negative' (true negatives) and 19 cases as 'Sick' (true positives). However, there were 19 instances where the model incorrectly predicted 'Sick' when the actual class was 'Negative' (false positives), and 2 instances where it predicted 'Negative' when the actual class was 'Sick' (false negatives). This performance demonstrates the model's strength in correctly identifying 'Negative' cases while highlighting the need for improvement in reducing false positives for 'Sick' predictions.

**Naive Bayes Classifier:** The model accurately identified 489 cases as 'Negative' (true negatives) and 25 cases as 'Sick' (true positives). However, it also had 21 instances where it incorrectly predicted 'Sick' instead of 'Negative' (false positives) and 13 instances where it predicted 'Negative' instead of 'Sick' (false negatives). While the model excels in correctly identifying 'Negative' cases, there is room for improvement to reduce false positives when predicting "Sick" cases and enhance overall precision.

**SVM Classifier:** The model effectively identified 509 cases as 'Negative' (true negatives) and 17 cases as 'Sick' (true positives). Nevertheless, it's important to note that there were 21 instances where the model falsely predicted 'Sick' when the actual class was 'Negative' (false positives), and 1 instance where it predicted 'Negative' when the actual class was 'Sick' (false negatives). While the model exhibits strong performance in correctly identifying 'Negative' cases, there is room for improvement to reduce false positives and enhance precision in predicting 'Sick' cases.

**Neural Network:** Out of 548 instances, the model correctly classified 496 as "negative" cases, indicative of a healthy thyroid status. However, it exhibited some difficulty in identifying "sick" cases, with 18 false-negative predictions.Conversely, the model correctly predicted 20 "sick" cases but also misclassified 14 "negative" cases as "sick".This highlights a tendency for the model to err on the side of caution, possibly leading to an increased number of false negatives. Further evaluation, including precision, recall, and F1-score analysis, is necessary for a comprehensive assessment of its predictive capabilities.

```
[1] "Accuracy of Random Forest:  0.9763"

[1] "Accuracy of XG Boost:  0.9818"
```

```
[1] "Accuracy of Logistic Regression:  0.9617"

[1] "Accuracy of Naive Bayes Classifier:  0.938"

[1] "Accuracy of SVM:  0.9599"

[1] "Accuracy of Neural Network:  0.9562"
```

The accuracy of different machine learning models provides valuable insights into their performance on the classification task at hand. Among the models evaluated, the **XG Boost model** stands out with an impressive accuracy of 98.18%. This exceptional accuracy underscores the model's capability to make accurate predictions, likely due to its ensemble approach and ability to handle complex datasets effectively.

Following closely behind is the **Randon Forest model**, which achieved an accuracy of 97.63%. Random Forest algorithm showcases its strength in improving model accuracy. While not surpassing XG Boost algorithm, it remains a highly competitive choice for accurate classification. (See appendix A)

In contrast, the **Logistic Regression model** achieved an accuracy of 96.17%, demonstrating strong performance. Logistic Regression, a linear model, showcases that simplicity can still lead to accurate results, making it a viable option when interpretability and ease of implementation are priorities. (See appendix C)

The **Naive Bayes Classifier** achieved an accuracy of 93.80%, showcasing its efficiency and simplicity. Despite slightly lower accuracy compared to the other specified models, Naive Bayes remains a suitable choice for quick classification tasks, particularly when dealing with limited data. (See appendix D)

The **Support Vector Machine (SVM) model** achieved an accuracy of 95.99%. SVMs are known for their robust classification capabilities and strong mathematical foundation. While performing well, it falls just short of the accuracy achieved by Random Forest and XG Boost models. (See appendix E)

Lastly, the **Neural Network Model** achieved an accuracy of 95.62%, indicating its ability to correctly classify thyroid disease cases. However, a more comprehensive evaluation should consider additional metrics such as precision and recall to gain a deeper understanding of its predictive performance. (See appendix F)

In conclusion, the choice of the most appropriate machine learning model should take into account factors such as the nature of the dataset, available computational resources, and the importance of interpretability. Random Forest and XG Boost models demonstrated the highest accuracy, making them top contenders for tasks requiring precise classification.

However, Logistic Regression, Naive Bayes, SVM and Neural Networks also showcased respectable performance and may be preferable under different circumstances.

## 5. Conclusions

In the pursuit of improving the early and accurate diagnosis of thyroid diseases, this investigation harnesses the power of machine learning. Leveraging the "MLDataR" package in R and drawing upon data from the UCI Machine Learning repository, this study delves into the multifaceted realm of thyroid health.

Most variables exhibit a typical bell-shaped curve, indicative of a normal distribution, suggesting that data points tend to cluster around the mean. However, TSH and T3 values defy this trend, displaying positive skewness, signaling potential health conditions associated with abnormal levels. The gender variable's influence on thyroid sickness is not strong, but the proportion within each category offers insights. Other variable pairs show weaker correlations, but their inclusion in the model may enhance accuracy. No significant relationship with thyroid class, warranting its exclusion.

The study employs various machine learning models, as depicted in Figure 4.1, with ROC curves showcasing their effectiveness in distinguishing thyroid classes. Notably, all models outperform the baseline, with Logistic Regression demonstrating relatively lower AUC. The XG Boost model excels in identifying "Negative" and "Sick" cases, boasting a 98.18% accuracy, followed closely by Random Forest Model at 97.63%.

Overall, these findings illuminate the potential of machine learning in thyroid disease diagnosis, with XG Boost Model emerging as a standout performer in accuracy, offering promising avenues for enhanced patient care.

## 6. References

[1] Thyroid-Stimulating Hormone (TSH) Levels, Article by Cleveland Clinic

[2] Thyroid Data Description by R Package Documentation

[3] Chi-Square test with mosaic plot for visualizations, Antoine Soetewey (2020)

[4] Thyroid Disease EDA, Classification and Ensembling, Kaggle, Elijah Rona (2021)

[5] Confusion matrix with cvms, Ludvig Renbo Olsen (2023)

## 7. Appendices

### APPERNDIX A: R CODES OF RANDOM FOREST CLASSIFIER

```
library(ranger)
## create the model
thyroid_ranger <- rand_forest(tree = 100, mode = 'classification') %>%
set_engine("ranger") %>%  #    use the random forest in ranger package
fit(ThryroidClass ~ ., data = thyroid_training)
```

### APPERNDIX B: R CODES OF XG BOOST CLASSIFIER

```
## create the model
thyroid_xgb <- boost_tree() %>%
set_engine(engine = "xgboost") %>%
set_mode("classification") %>%
fit(ThryroidClass ~ ., data = thyroid_training)
```

### APPERNDIX C: R CODES OF LOGISTIC REGRESSION CLASSIFIER

```
## create the model
thyroid_log <- logistic_reg() %>%
set_engine(engine = "glm") %>%
set_mode("classification") %>%
fit(ThryroidClass ~ ., data = thyroid_training)
```

## APPERNDIX D: R CODES OF NAIVE BAYES CLASSIFIER

```
library(e1071)

## create the model

thyroid_NB <- naiveBayes(ThryroidClass ~ ., data = thyroid_training)

## Prediction for the test test

Predict_thy_NB <- predict(thyroid_NB, thyroid_testing)
```

## APPERNDIX E: R CODES OF SVM CLASSIFIER

```
library(kernlab)

## create the model

thyroid_svm <- svm_rbf(cost = 0.5)

set_engine("kernlab") %>%

set_mode("classification") %>%

fit(ThryroidClass ~ ., data = thyroid_training)
```

## APPERNDIX F: R CODES OF NEURAL NETWORK

```
library(nnet)

## create the model

thyroid_NN <- nnet(ThryroidClass ~ ., data = thyroid_training, size = 5, lino
ut = FALSE)

Predict_thy_NN <- predict(thyroid_NN, newdata = thyroid_testing, type = "clas
s")
```

***To Access Full Source Code:***

*AyomiUpeksha/Thyroid-Class-Classification: Machine Learning for Thyroid Disease Prediction 🤖 🩺; A robust predictive model (github.com)*