Predictive Modeling of Thyroid Function

Toni Crnjak

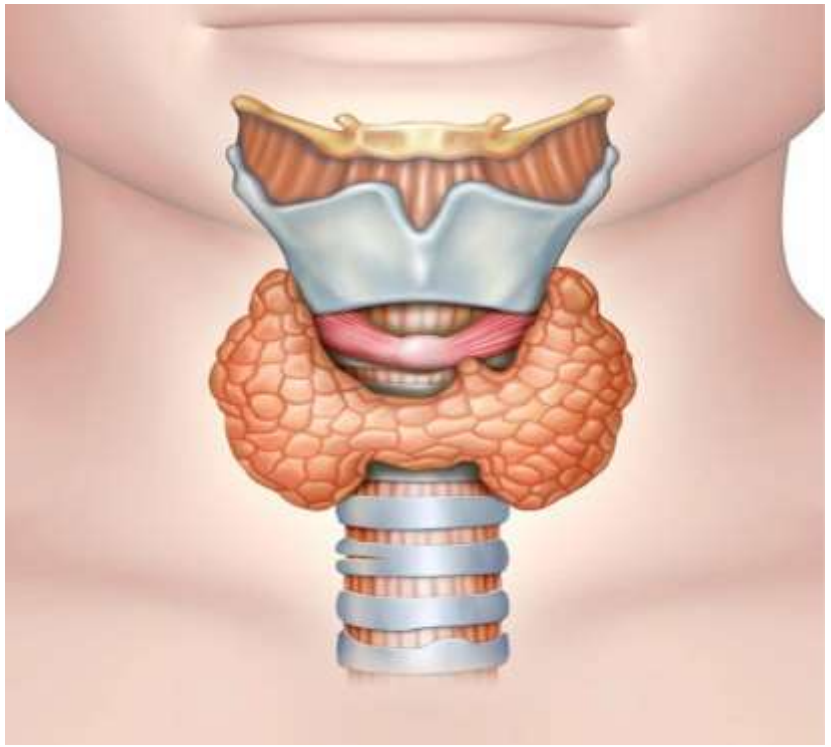
Mounika Kaluvakuntla

IA650- Data Mining Project

04 December 2025

Introduction

- Thyroid regulates metabolism, energy, and growth
- Key hormones: Triiodothyronine(T3), Thyroxine(T4), and Thyroid stimulating hormone(TSH)
- Hormonal imbalance → Hypo/Hyperthyroidism
- The dataset contains 1848 records and 9 variables, including demographic details and biochemical test results related to thyroid function.
  - 1848 Rows
  - 9 Variables
  - 1 Target Variable- Thyroid Range
- Source: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023



https://teachmeanatomy.info/neck/viscera/thyroid-gland/

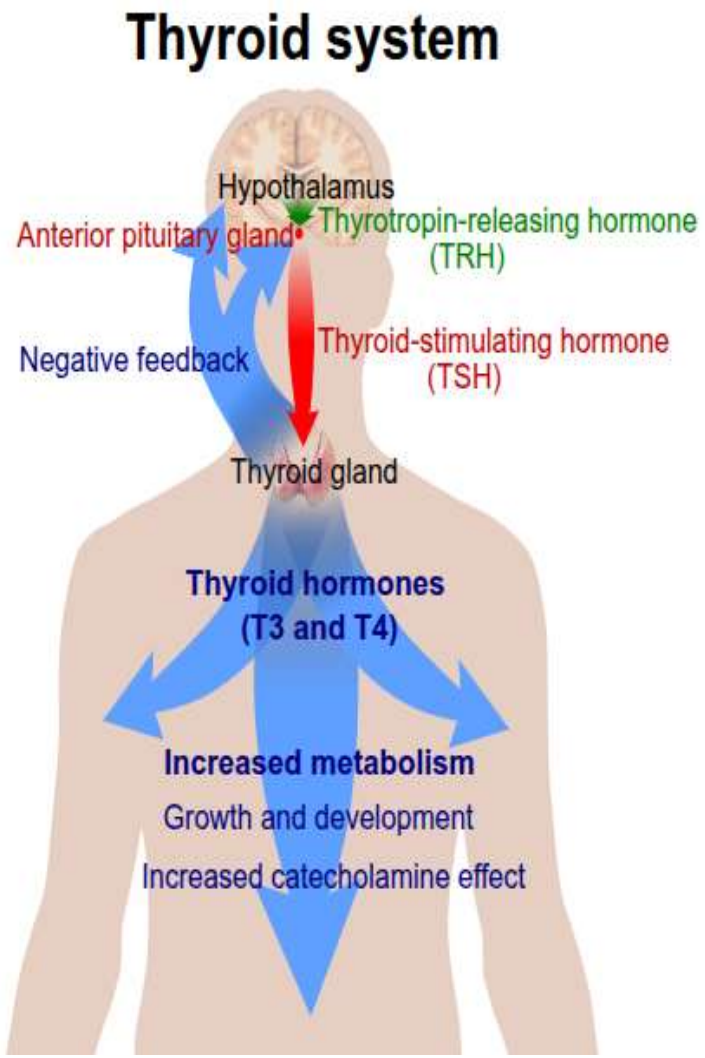
Hormone Definitions and Functions

- **Triiodothyronine (T3):** Active hormone controlling metabolism
- **Thyroxine (T4):** The main hormone secreted by the thyroid gland
- **Thyroid Stimulating Hormone (TSH):** Pituitary control hormone
- **Thyroid Range:** Target variable (Low, Normal, High)
- **Normal Ranges :-**

  T3: Hypothyroid < (2.2, 3.1) < Hyperthyroid

  T4: Hypothyroid < (0.6, 1.45) < Hyperthyroid

  TSH: Hyperthyroid < (1, 4) < Hypothyroid



https://en.wikipedia.org/wiki/Thyroid

Objective

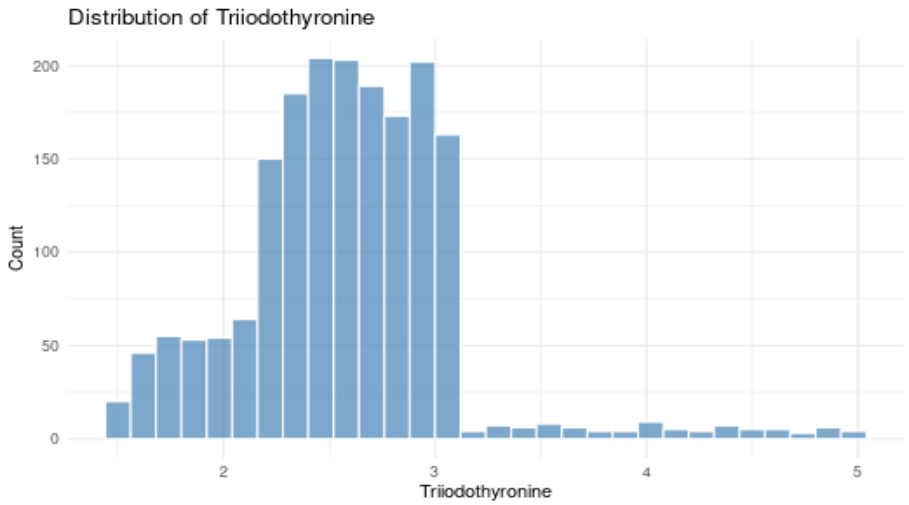
- Identify key patterns and relationships among the thyroid-related variables
- Develop predictive model
- Enable early diagnosis of thyroid abnormalities

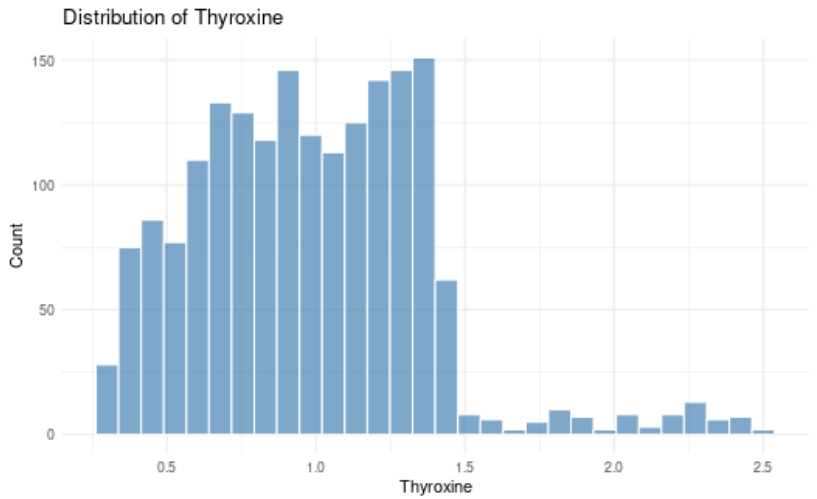
https://www.betheshyft.com/community/thread/NDk5/symptoms-of-thyroiddisorder/

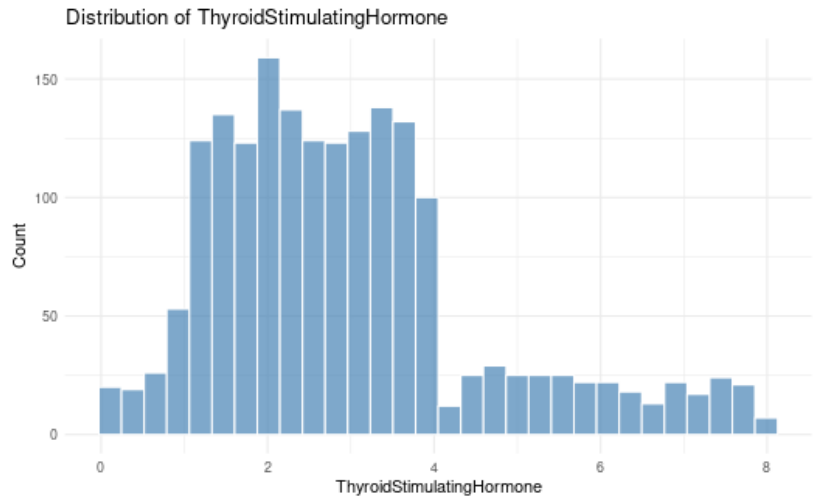
Variables

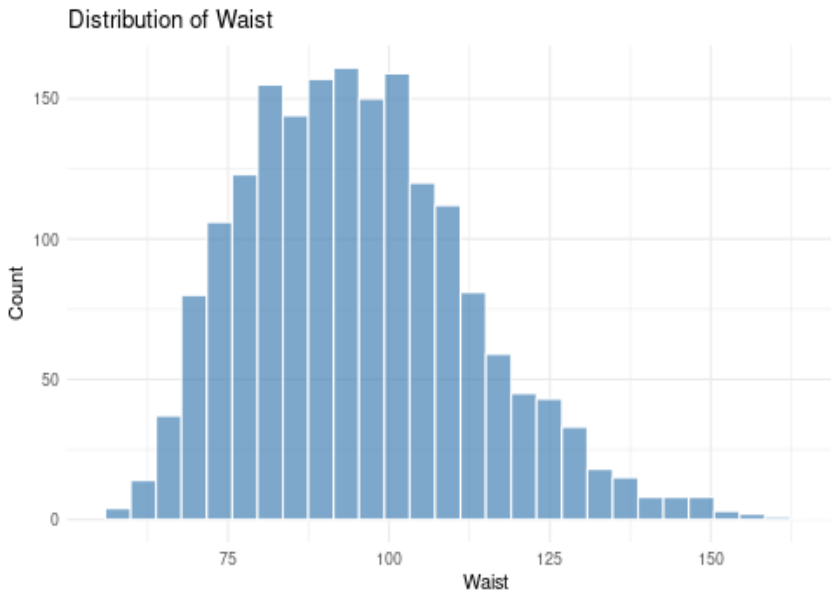
Variable Name	Description	Type of Variable	Measurement Unit
Gender	Biological sex of the patient (1 = Male, 2 = Female)	Categorical	-
Age	Age of the patient	Numerical	Years
Waist	Waist circumference measurement	Numerical	cm
Height	Height of the patient	Numerical	cm
BMI	Body Mass Index calculated from height and weight	Numerical	kg/m^2
Triiodothyronine (T3)	triiodothyronine hormone level	Numerical	pg/mL
Thyroxine(T4)	thyroxine hormone level	Numerical	ng/dL
Thyroid stimulating hormone(TSH)	Thyroid Stimulating Hormone concentration	Numerical	µIU/mL
Thyroid Range	Categorized range of Thyroid (1 = Low, 2 = Normal, 3 = High)	Categorical	-

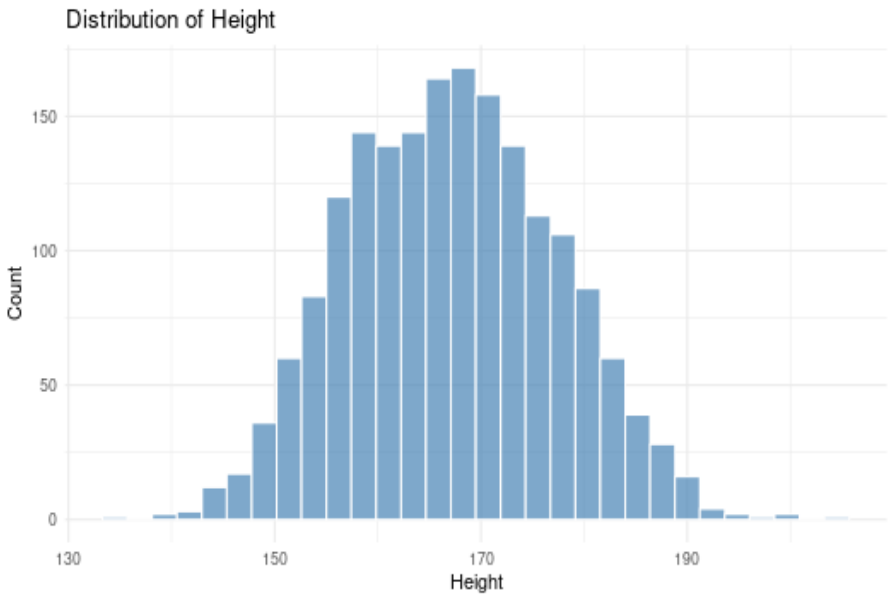
Distribution of Numeric Variables

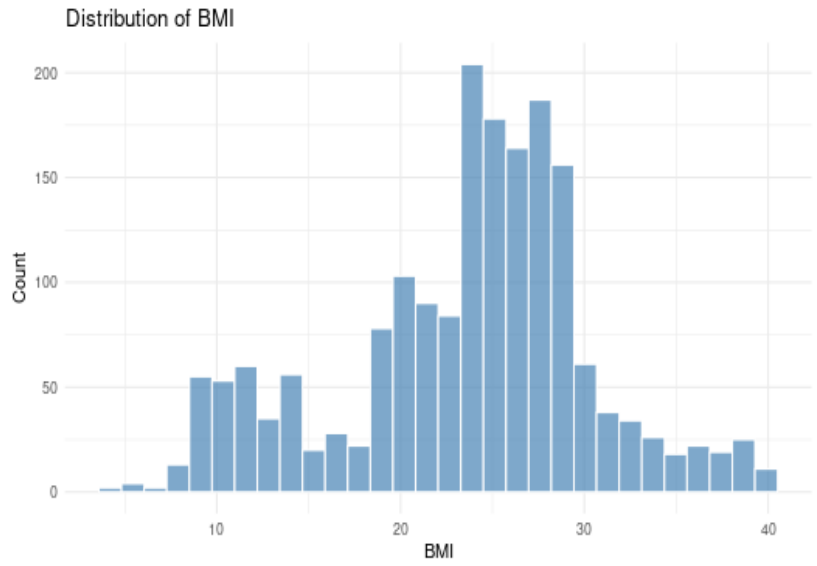




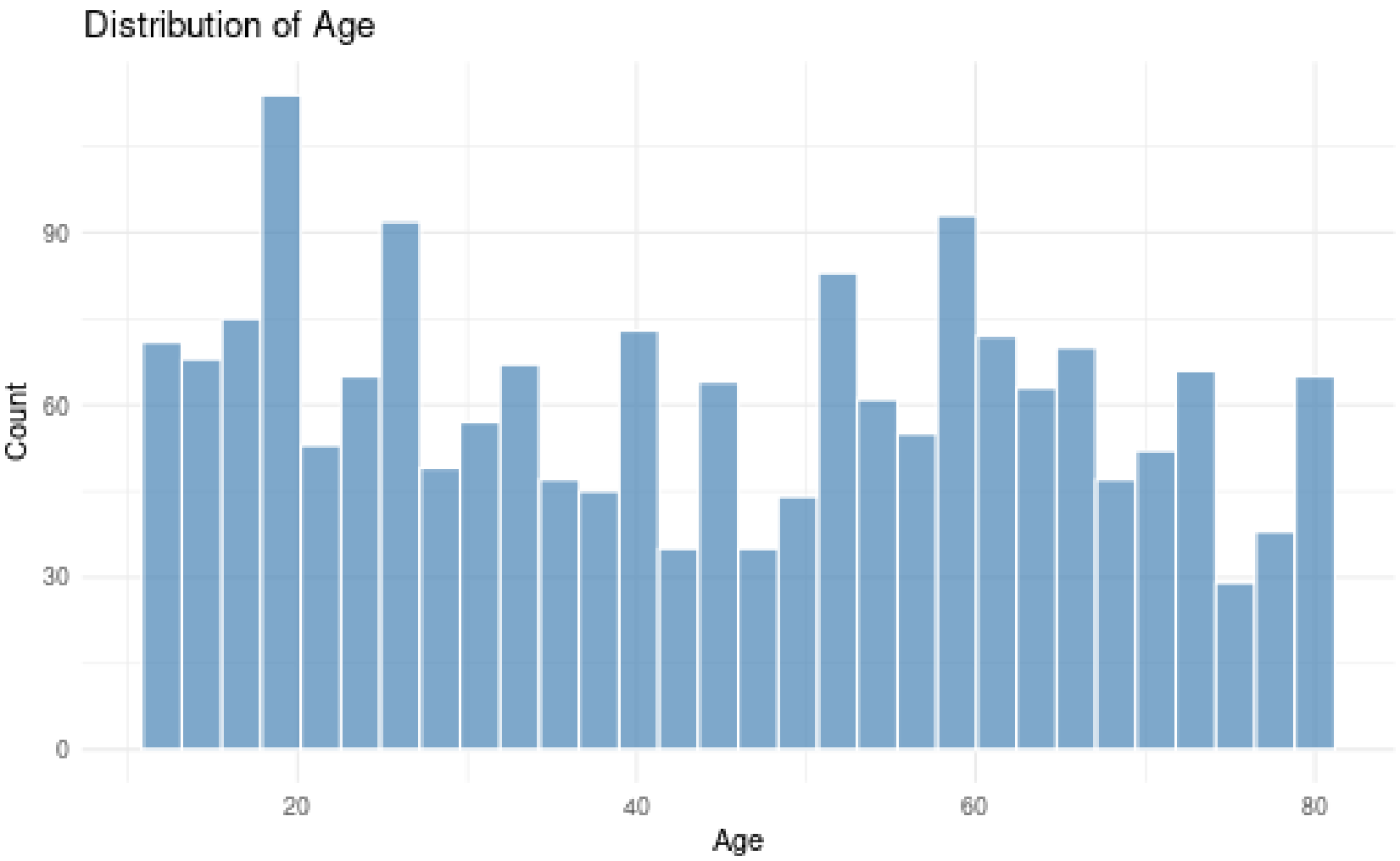






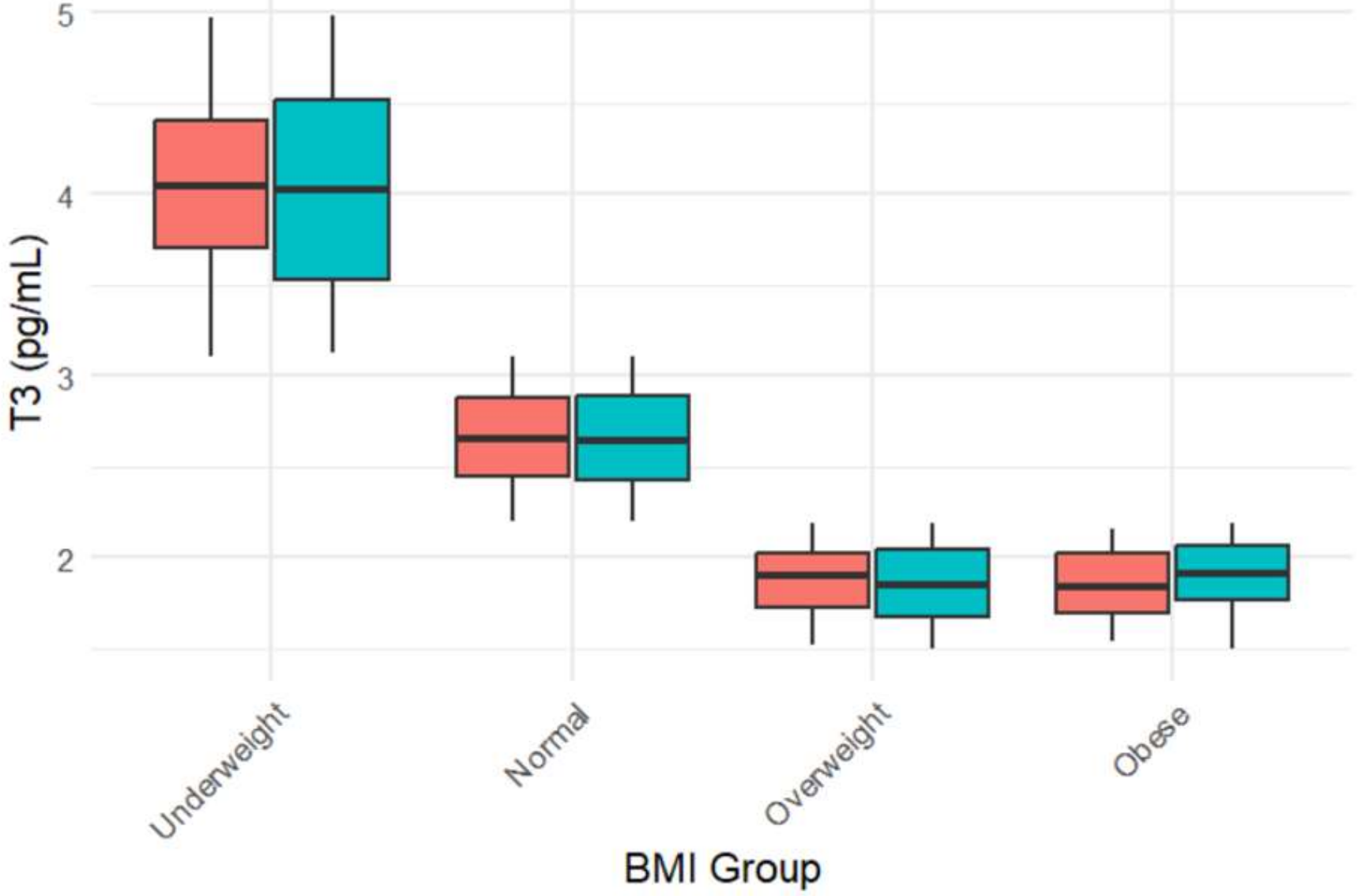


Distribution of Numeric Variables



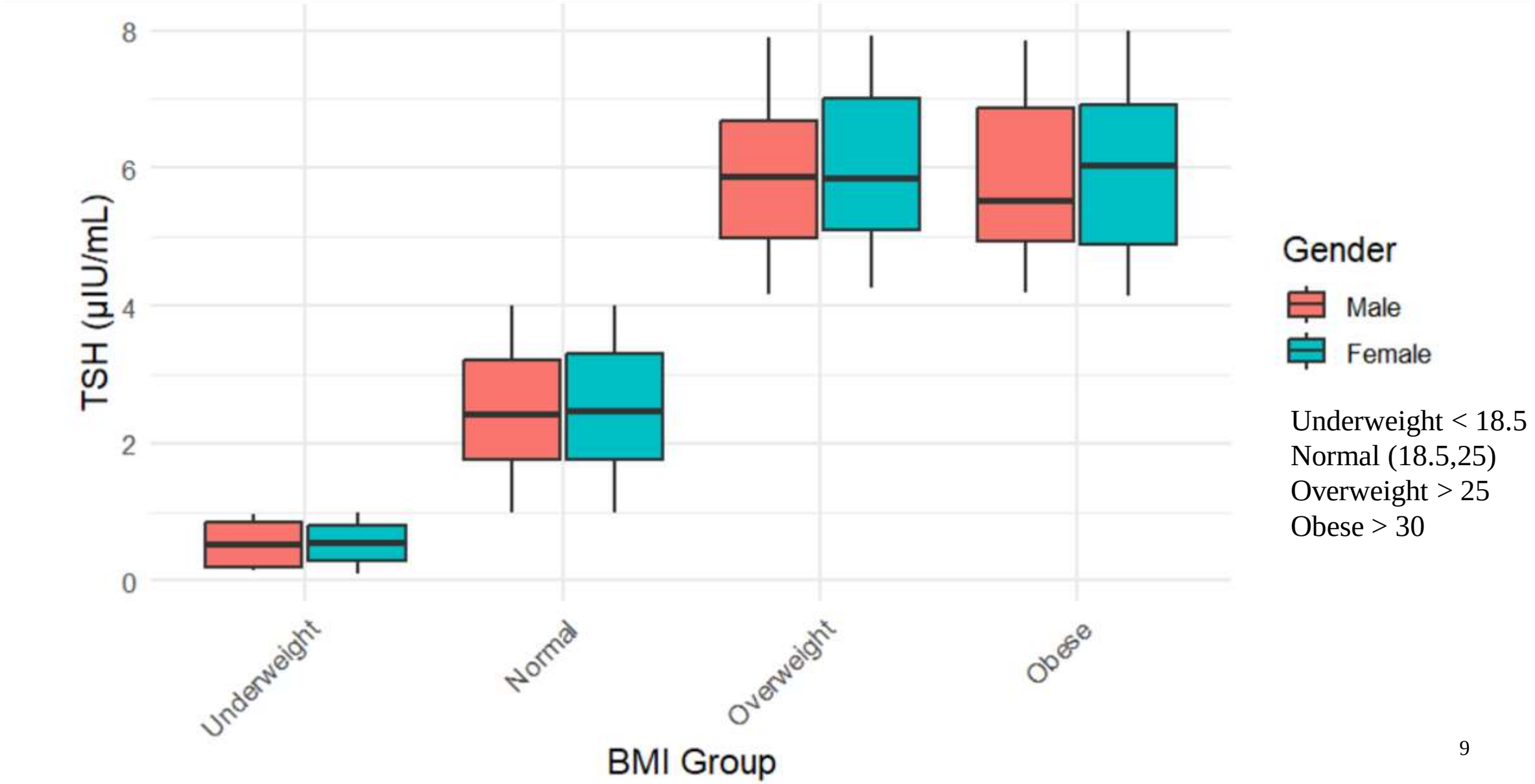
- The numeric variables show skewed and non-normal distributions, which is typical in real-world health and hormone data.
- The variability seen across age, BMI, hormones, and anthropometrics reflects natural population differences rather than noise.
- Even though the data are not normally distributed, they remain highly reliable because they come from standardized CDC/NCHS collection protocols.

Boxplot of Triiodothyronine by BMI group (colored by Gender)



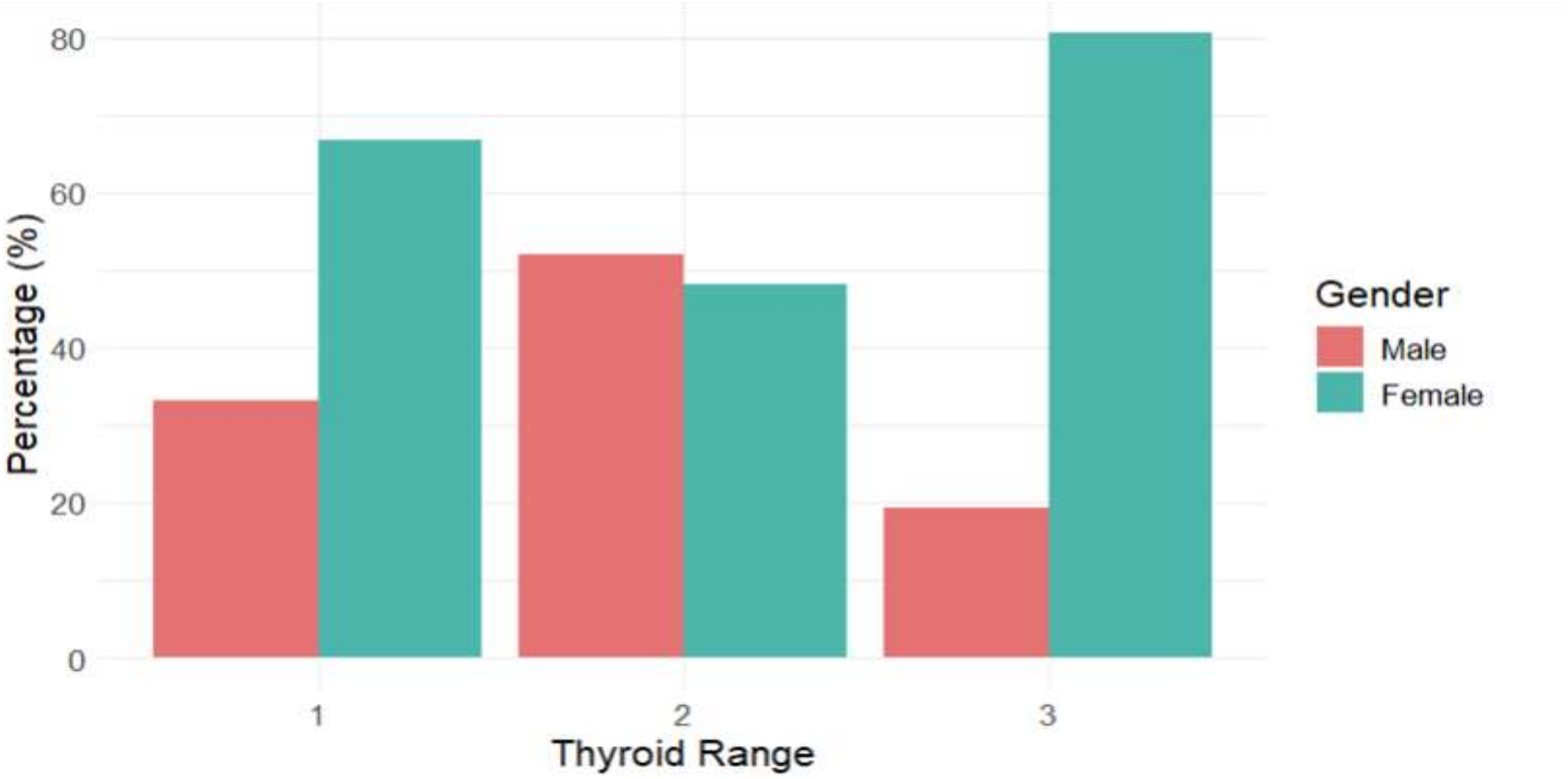
Underweight < 18.5
Normal (18.5,25)
Overweight > 25
Obese > 30

Boxplot of TSH by BMI group (colored by Gender)

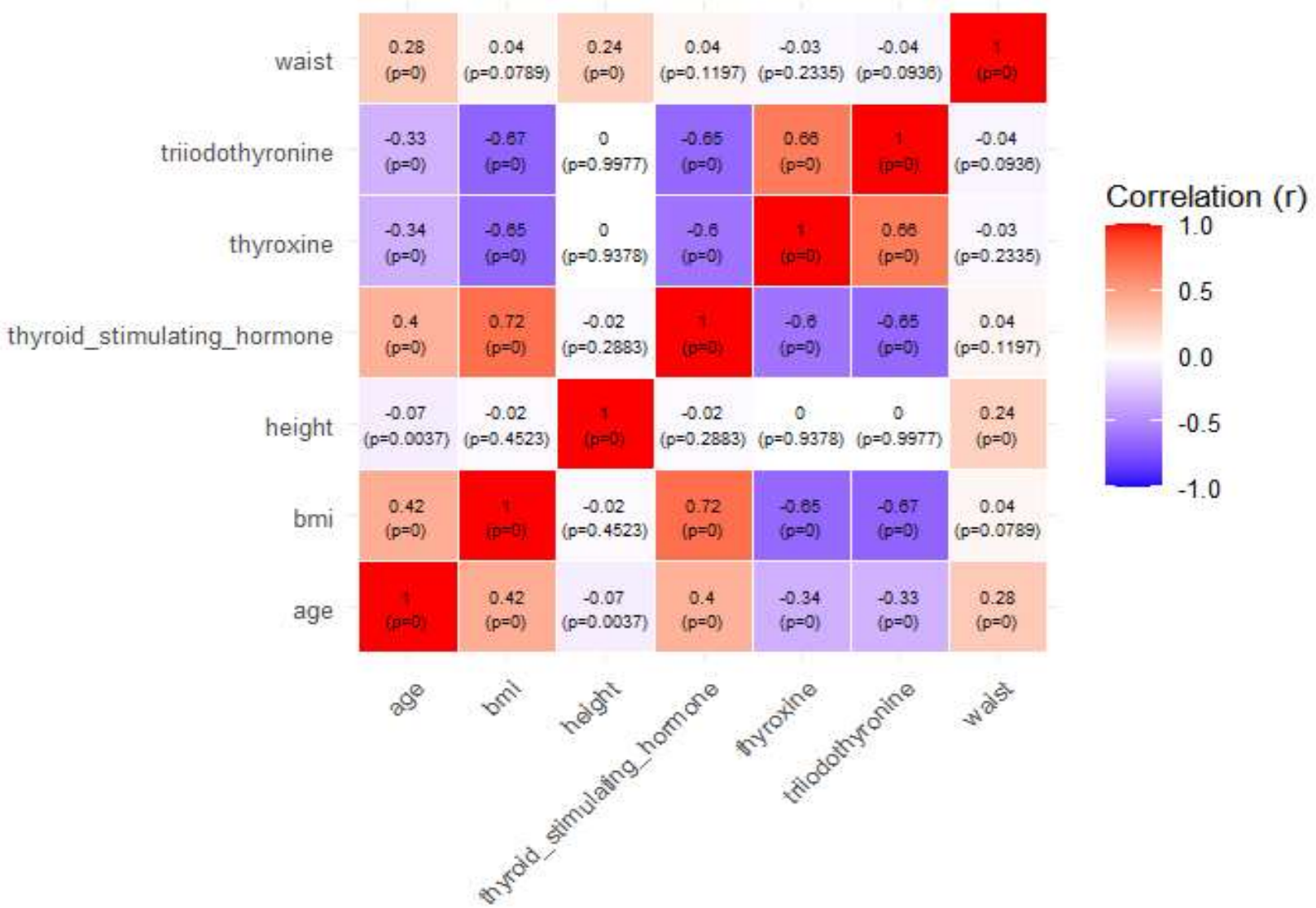


Underweight < 18.5
Normal (18.5,25)
Overweight > 25
Obese > 30

Gender Distribution Across Thyroid Ranges



Correlation Matrix



- Strong positive correlation between T3 and T4, and strong negative correlation with TSH
- Moderate associations between BMI, age, and thyroid hormone levels
- Patterns align with established physiological relationships, supporting model reliability

Chi-Squared for Categorical Variables

H_0: Gender is independent of the Thyroid Range.

H_1: Gender is associated with the Thyroid Range.

Feature	Statistic	P-value	Result
Gender	64.341	<0.05	Significant

The chi-square test shows a highly significance between Gender and Triiodothyronine Range ($\chi^2 = 64.34$, $df = 2$, $p < 0.0001$)

Normality Test - Shapiro-wilk test

H_0: There is no difference in the distribution of the variable between Thyroid Ranges.

H_1: There is significant difference in the distribution of the variable between Thyroid Ranges.

Feature	Statistic	p-Value	Result
Gender	0.6355	<0.05	Not Normal
Age	0.943	<0.05	Not Normal
Waist	0.978	<0.05	Not Normal
Height	0.996	<0.05	Not Normal
BMI	0.965	<0.05	Not Normal
Triiodothyronine	0.918	<0.05	Not Normal
Thyroxine	0.9505	<0.05	Not Normal
Thyroid Stimulating Hormone	0.923	<0.05	Not Normal
Thyroid Range	0.608	<0.05	Not Normal

Kruskal Wallis Test for Numerical Variables

H_0: There is no difference in the distribution of the variable between Thyroid Ranges.

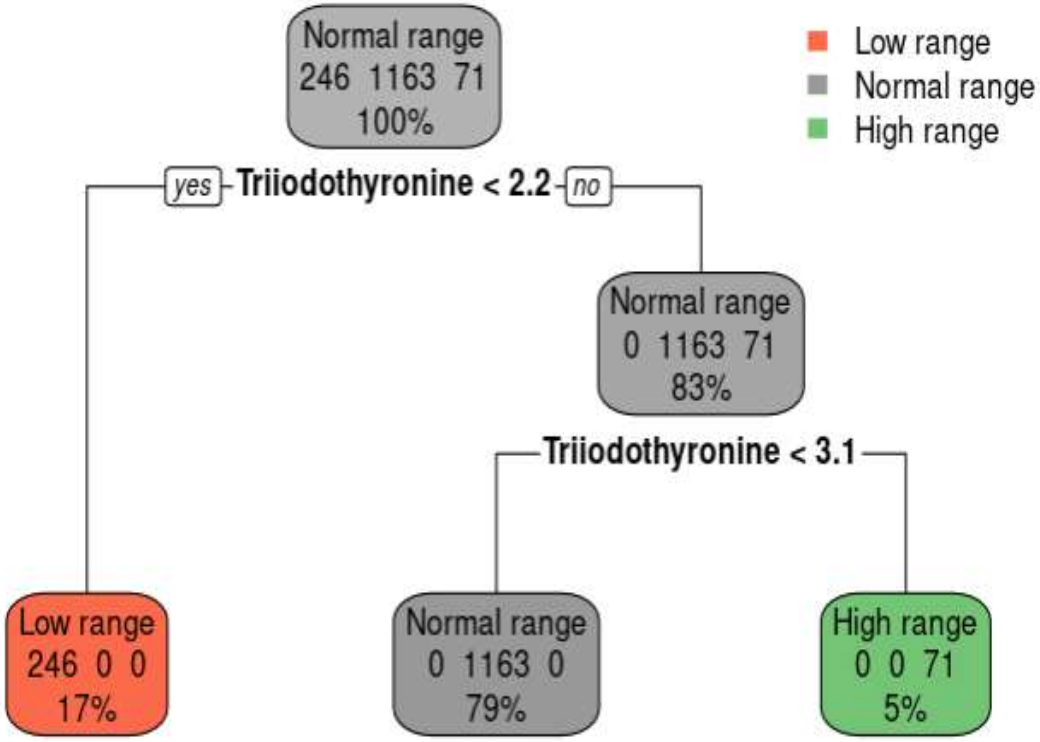
H_1: There is significant difference in the distribution of the variable between Thyroid Ranges.

Feature	Statistic	p-Value	Result
Age	401.19	<0.05	Significant
Waist	6.09	<0.05	Significant
Height	4.09	>0.05	Not Significant
BMI	409.71	<0.05	Significant
Triiodothyronine	940.609	<0.05	Significant
Thyroxine	940.69	<0.05	Significant
Thyroid Stimulating Hormone	940.59	<0.05	Significant

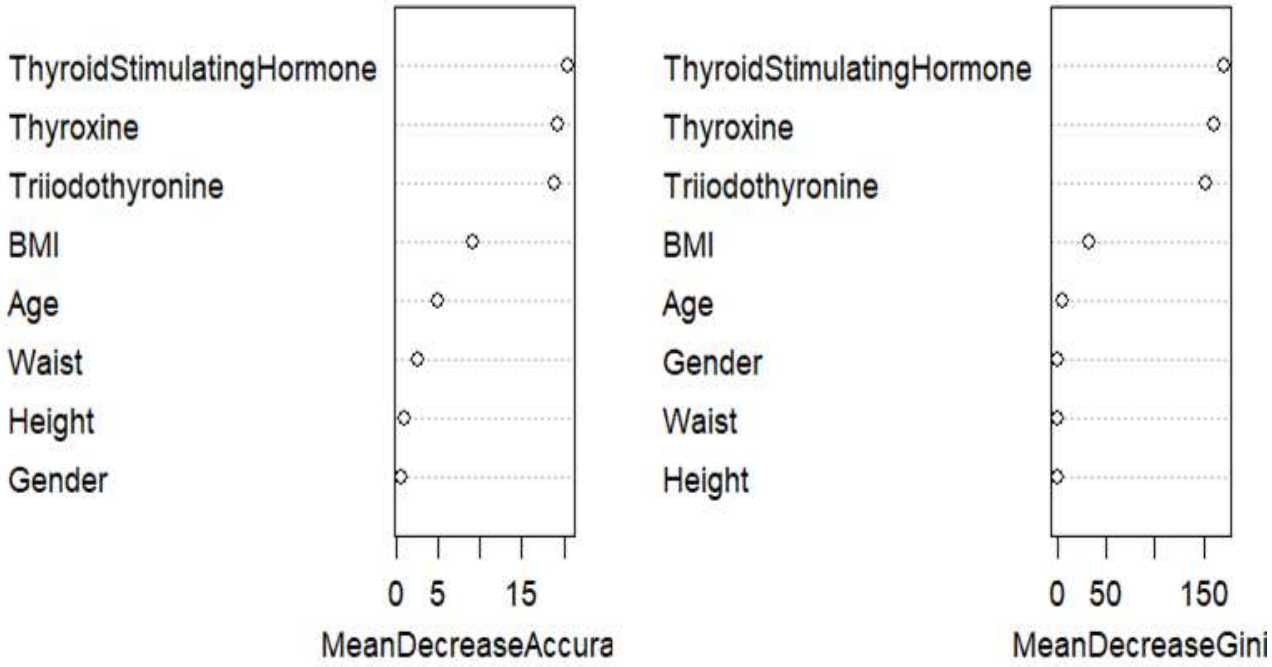
Logistic Regression

Predictor	Coefficient	CI (2.5%)	CI (97.5%)
Thyroid Range	152.417	3.78e^-15	1.25e^133
Gender	-15.224	4.72e^-150	1.09e^-109
Age	-0.69	1.17e^-65	2.82e^+58
Waist	0.107	2.28e^-321	Inf
Height	-1.35	4.87e^-63	8.90e^+61
BMI	-0.77	4.97e^-63	7.29e^+61
Triiodothyronine (T3)	95.74	2.25e^-83	1.04e^+83
Thyroxine (T4)	129.22	3.34e^-125	3.27e^+124
TSH (Thyroid Stim. Hormone)	-37.104	8.96e^-38	3.33e^+36

Decision Tree

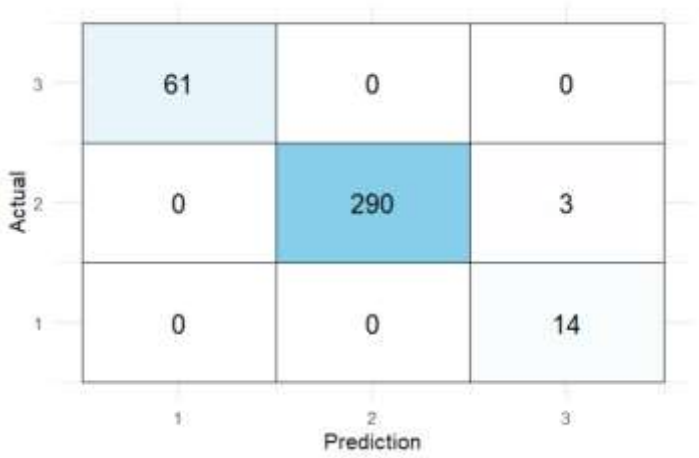


Random Forest

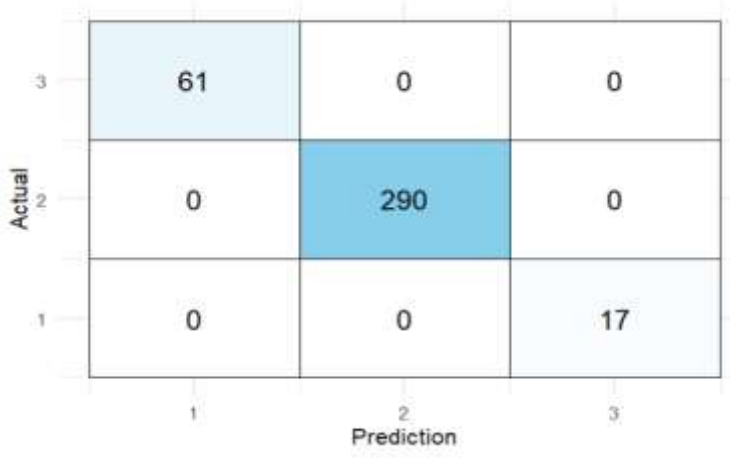


Confusion Matrices

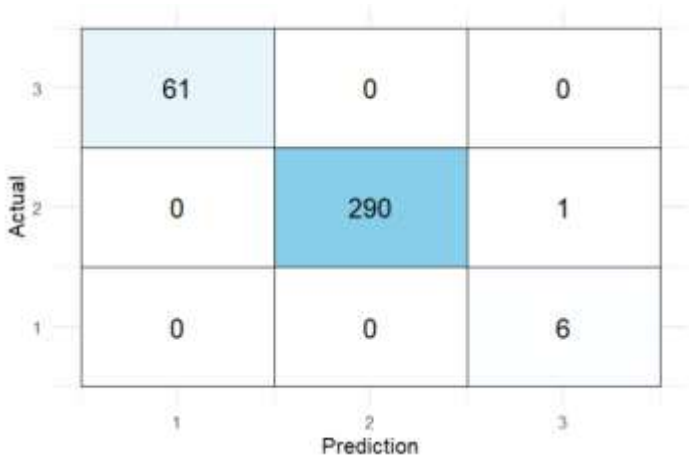
Logistic Regression



Decision Tree



Random Forest



Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.992	0.941	0.997	0.966
Decision Tree	1	1	1	1
Random Forest	0.997	0.9524	0.9989	0.9738

Conclusion

- Females have high tendency to get thyroid disorders → gender-related differences
- Older age & higher BMI groups → more abnormal thyroid values
- Strong positive correlation between T3 and T4, and negative correlation between TSH with T3/T4, aligns with known medical physiology.
- Among all models tested, the Decision Tree achieved the best overall performance, indicating strong rule-based relationships in the data.

Future Scope

- Collect more clinical parameters (diet, medication history, symptoms, family history) for richer modeling and improved diagnosis which is balanced.
- Apply advanced machine learning models such as Gradient Boosting, XGBoost, or Neural Networks.
- Deploy as a clinical decision support system.

Acknowledgement

Professor Sumona Mondal

Professor Naveen Ramachandra Reddy

THANK YOU