**FINAL PROJECT**

Rümeysa Durdağ   
*Middle East Technical University*Ankara, Turkey  
rumeysa.durdag@metu.edu.tr

**Abstract— The thyroid hormone is produced by thyroid gland. This hormone regulates the body’s metabolism. This study presents the causative agents of Thyroid disease and predicts the condition of these patients with different machine learning methods such as Support Vector Machine, Neural Network, Xgboost, Random Forest. Before generating the prediction, the data is cleaned to remove any obvious errors. The models are executed after obtaining the clean dataset. R-Studio is used to perform all procedures.**

Keywords—SVM, Artificial Neural Network, Random Forest, XGBoost

1. INTRODUCTION

To control the body's metabolism, the thyroid gland produces thyroid hormones. The thyroid gland makes the hormones triiodothyronine (T3) and levothyroxine (T4). These hormones are crucial in turning food into energy, which in turn controls body temperature, heart rate, and even brain function. When the thyroid gland produces more hormones than the body needs, hyperthyroidism results. However, hypothyroidism happens when the gland produces fewer hormones than what the body needs. Deaths may result from severe hyperthyroidism and end-stage hypothyroidism. It may result in maternal and fetal health issues during pregnancy.

Patients can be spared if these fatal illnesses are caught early enough, though. According to this viewpoint, some machine learning techniques can be used to address this type of problem.

This study estimates whether patients have thyroid illness or not using a variety of regression models such as Support Vector Machine, Neural Network, Xgboost, and Random Forest.

1. LITERATURE REVIEW

Several studies have investigated the application of machine learning algorithms for detecting thyroid disease using various datasets and methodologies. Zhang et al. (2019)[1] focused on the classification of thyroid nodules as benign or malignant. They utilized features derived from medical images and clinical data to train models, and their results indicated that Random Forest achieved high accuracy in identifying malignant nodules. This finding highlights the potential of machine learning techniques in assisting with the accurate diagnosis of thyroid cancer.

In another study by Li et al. (2020)[2], the authors aimed to differentiate between different types of thyroid nodules. They employed Deep Neural Networks and Random Forest models, utilizing features extracted from ultrasound images and clinical data.

Their findings demonstrated that machine learning models can effectively classify and distinguish between different types of thyroid nodules, with Deep Neural Networks outperforming Random Forest. This research showcases the utility of machine learning approaches in improving the diagnosis and management of thyroid diseases, offering valuable insights into potential advancements in the field.

1. METHODOLOGY
2. *Dataset*

The dataset contains the thyroid disease records collected and supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. 1987. The data set includes 30 variables and 3772 observations. It has 6064 missing values. The following section of the analysis will provide information on the missing observation. Below is a description of the variable.

1. **age** - age of the patient **(int)**
2. **sex** - sex patient identifies **(str)**
3. **on\_thyroxine** - whether patient is on thyroxine **(bool)**
4. **query on thyroxine** - \*whether patient is on thyroxine **(bool)**
5. **on antithyroid meds** - whether patient is on antithyroid meds **(bool)**
6. **sick** - whether patient is sick **(bool)**
7. **pregnant** - whether patient is pregnant **(bool)**
8. **thyroid\_surgery** - whether patient has undergone thyroid surgery **(bool)**
9. **I131\_treatment** - whether patient is undergoing I131 treatment **(bool)**
10. **query\_hypothyroid** - whether patient believes they have hypothyroid **(bool)**
11. **query\_hyperthyroid** - whether patient believes they have hyperthyroid **(bool)**
12. **lithium** - whether patient \* lithium **(bool)**
13. **goitre** - whether patient has goitre **(bool)**
14. **tumor** - whether patient has tumor **(bool)**
15. **hypopituitary** - whether patient \* hyperpituitary gland **(float)**
16. **psych** - whether patient \* psych **(bool)**
17. **TSH\_measured** - whether TSH was measured in the blood **(bool)**
18. **TSH** - TSH level in blood from lab work **(float)**
19. **T3\_measured** - whether T3 was measured in the blood **(bool)**
20. **T3** - T3 level in blood from lab work **(float)**
21. **TT4\_measured** - whether TT4 was measured in the blood **(bool)**
22. **TT4** - TT4 level in blood from lab work **(float)**
23. **T4U\_measured** - whether T4U was measured in the blood **(bool)**
24. **T4U** - T4U level in blood from lab work **(float)**
25. **FTI\_measured** - whether FTI was measured in the blood **(bool)**
26. **FTI** - FTI level in blood from lab work **(float)**
27. **TBG\_measured** - whether TBG was measured in the blood **(bool)**
28. **TBG** - TBG level in blood from lab work **(float)**
29. **referral\_source** - **(str)**
30. **class (str) (RESPONSE VARIABLE)**
31. *Descriptive Statistics*

Descriptive statistics table are shown below. It was obtained first, since it gives an insight into the data set at the beginning of the exploration. Besides, these values help to create research questions in the next steps of the analysis.

A descriptive summary is attached in Table 1 for numerical attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column1** | **age** | **TSH** | **T3** |
| Min. | 1 | 0.005 | 0.050 |
| 1st Qu. | 36.0 | 0.500 | 1.600 |
| Median | 54.0 | 1.400 | 2.000 |
| Mean | 51.63 | 5.087 | 2.014 |
| 3rd Qu. | 67.0 | 2.700 | 2.400 |
| Max. | 94.0 | 530.00 | 10.600 |
| NA's | 2 | 369 | 769 |

*Table 1 Descriptive Statistical Summary of Numerical Data*

|  |  |  |  |
| --- | --- | --- | --- |
| **Column1** | **TT4** | **T4U** | **FTI** |
| Min. | 2 | 0.250 | 2 |
| 1st Qu. | 88.0 | 0.880 | 93.0 |
| Median | 103.0 | 0.980 | 107.0 |
| Mean | 108.3 | 0.995 | 110.5 |
| 3rd Qu. | 124.0 | 1.080 | 124.0 |
| Max. | 430.0 | 2.320 | 395.0 |
| NA's | 231 | 387 | 385 |

*Table 2 Descriptive Statistical Summary of Numerical Data*

The continuous variables in the data set include this table. The minimum value, first quantile, median, third quantile, and maximum value of the variables are all represented by the Tukey's five number summary statistics. Additionally, the mean value of the given variables and the amount of missing values for each variable are both shown in the table.

1. *Exploratory and Confirmatory Analysis*

In this section, a few research questions have been asked and using the appropriate statistical techniques, answers are provided to address the problems in question. As was already said, these statistics provide a broad understanding of the variables and allow us to further the investigation.

*C.1 Are there any significant differences between Class and TSH?*

**

*Figure 1 Boxplot of Age, by Thyroid Sickness*

We have a hypothesis based on the graph that suggests older individuals may have a higher risk of thyroid illness. Because of this, we performed a T test (with in mind that the age variable in our sample has a normal distribution). T-test results revealed that there was a difference; the mean age of healthy individuals was 50.82, while that of sick individuals was 63.92, indicating that older individuals had a larger chance of developing thyroid disease.

Overall, based on the t-test results, there is strong evidence to suggest that there is a significant difference in mean ages between the "negative" and "sick" groups.

*C.2 Are there any significant correlations between different thyroid hormone levels in the dataset?*

****

*Figure 2 Boxplot of Age, by Thyroid Sickness*

Most of the variables are associated in figure 7 above, although not always in a very strong way. The Pearson's correlation coefficient was estimated to be 0.793, indicating a strong linear relationship between the two variables. The t-value of 75.778 with a corresponding p-value of less than 2.2e-16 suggests that this correlation is statistically significant.

The 95% confidence interval for the correlation coefficient ranges from 0.780 to 0.805, indicating that we can be highly confident that the true correlation in the population falls within this interval. The alternative hypothesis states that the true correlation is not equal to zero, which is supported by the significant p-value. Therefore, we can conclude that there is a strong positive correlation between TT4 and FTI.

1. *Missingness*

The dataset contains missing values. Missing values are visualized in the figure below. Missing data are loaded with the "mice" package in the R program.



*Figure3: Number of Missing Observations per Variable*

In Figure 4, it is evident that the missing values appear randomly throughout the dataset and are not dependent on one another.

## Modelling

The data set may be used to build a model that controls methodological correction and predicts the condition of patients after missing imputation. We then divide the data into train and test sets, and from each set we construct a model. Cross-Validation is the name of this process. The reason for this is that we want to see if our model still makes sense after taking into account new data on the variables.

In order to build a model, we thus utilize 80% of the data set. In addition, we accept 20% of the data set, as test data.

1. *Support Vector Machine*

A supervised learning method called the support vector machine can be applied to classification and regression problems.

|  |  |
| --- | --- |
|  | **Reference** |
| Prediction No Yes | |
| No 2822 96 | |
| Yes 11 89 | |

*Table 1 : Train*

|  |
| --- |
| **Reference** |
| Prediction No Yes |
| No 705 27 |
| Yes 3 19 |

|  | **Accuracy** | **NIR** |
| --- | --- | --- |
| Train | 0.9645 | 0.9387 |
| Test | 0.9602 | 0.939 |

*Table 2 : Test*

The accuracy of the model is calculated as 0.9645, indicating that it correctly classified 96.45% of the instances. The model's accuracy is significantly higher than the no information rate (NIR) of 0.9387, which suggests that the model performs better than random chance.

An overall measure of categorization performance, the balanced accuracy, is 73.86%. According to the designation of the model, the "No" class is regarded as the positive class.

1. *Artificial Neural Network*

Another supervised learning approach that can be applied to classification and regression problems is the artificial neural network.

|  |  |
| --- | --- |
|  | **Reference** |
| Prediction 0 1 | |
| 0 2806 26 | |
| 1 27 159 | |

*Table 1 : Train*

|  |  |
| --- | --- |
|  | **Reference** |
| Prediction 0 1 | |
| 0 700 9 | |
| 1 8 37 | |

|  | **Accuracy** | **NIR** |
| --- | --- | --- |
| Train | 0.9824 | 0.9387 |
| Test | 0.9775 | 0.939 |

The artificial neural network (ANN) model achieved an accuracy of 98.24% in predicting the target variable.

the NIR is 93.87%. When comparing the accuracy of the neural network model (98.24%) to the NIR, it indicates that the model significantly outperforms the baseline of predicting the most frequent class.

*Table 2 : Test*

The model shows strong performance with high accuracy and balanced accuracy. It achieves a good balance between sensitivity and specificity, indicating its ability to correctly classify both positive and negative cases.

1. *Random Forests*

Another supervised learning method based on tree algorithms that can be used for both regression and classification issues is random forests.

|  |  |
| --- | --- |
|  | Reference |
| Prediction 0 1 | |
| 0 2832 0 | |
| 1 1 185 | |

*Table 1 : Train*

|  |  |
| --- | --- |
|  | Reference |
| Prediction 0 1 | |
| 0 703 10 | |
| 1 5 36 | |

*Table 2 : Test*

|  | **Accuracy** | **NIR** |
| --- | --- | --- |
| Train | 0.9997 | 0.9387 |
| Test | 0.9801 | 0.939 |

The random forest model exhibits exceptional performance, with near-perfect accuracy and high values for various evaluation metrics. It demonstrates its ability to accurately classify instances, with an extremely low error rate and a high degree of confidence in its predictions.

Overall, the model exhibits good performance, with a high accuracy rate and satisfactory values for various evaluation metrics. It demonstrates its ability to accurately classify instances, particularly positive cases, but may have some room for improvement in correctly identifying negative cases.

1. *XgBoost*

Xgboost is a machine learning method based on tree algorithm and can be used for both regression and classification problems.

|  |  |
| --- | --- |
| Reference | Reference |
| Prediction 0 1 | |
| 0 701 7 | |
| 1 5 41 | |

|  | **Accuracy** | **NIR** |
| --- | --- | --- |
| Train | 0.9841 | 0.9363 |
| Test | 0.9841 | 0.9363 |

1. *Perfomance Comparison and Results*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NIR | ACC | SN | SP |
| SVM | 0.939 | 0.9602 | 0.9958 | 0.4130 |
| ANN | 0.939 | 0.9775 | 0.9887 | 0.8043 |
| RF | 0.939 | 0.9801 | 0.9929 | 0.7826 |
| XGBoost | 0.9363 | 0.9841 | 0.9929 | 0.8542 |

When comparing all of the values in table, the best model to predict Thyroid disease occurence is XgBoost followed by RandomForest, ANN and SVM in this order .

# CONCLUSION

In this work, the analysis of the Thyroid Sickness Determination dataset revealed significant associations between variables related to patient

demographics and thyroid hormone levels. Machine learning models, such as Support Vector Machine, Artificial Neural Network, and Random Forest, XgBoost, demonstrated high accuracies in predicting thyroid sickness. These findings highlight the potential of machine learning techniques in aiding the early detection and diagnosis of thyroid disease. However, further validation and refinement of these models are needed, considering the limitations of the dataset and potential biases. Continued research in this field will enhance our understanding and improve diagnostic capabilities for thyroid sickness.

# REFERENCES

[1] Zhang, Y., Zhao, Y., Wang, X., Wang, R., & Zhu, Y. (2019). A Comparative Study of Machine Learning Algorithms in the Classification of Thyroid Nodules Using Ultrasound Images. Journal of Medical Systems, 43(7), 198.

[2] Li, Y., Huang, Z., Huang, Q., Zhou, Q., & Qin, B. (2020). Deep Learning and Machine Learning in Differentiating Thyroid Nodules: A Pilot Study. Journal of Ultrasound in Medicine, 39(12), 2529-2537.