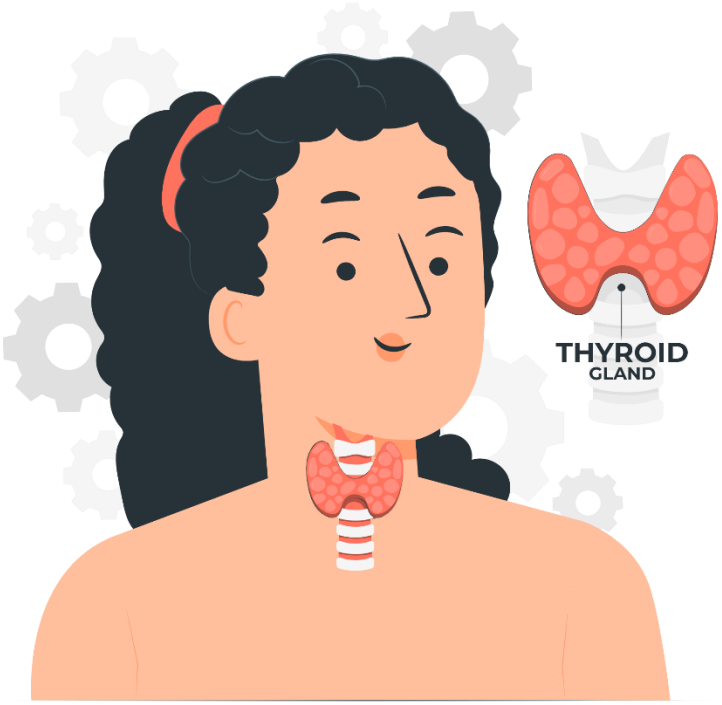
TITLE

EXAMINING THE USE OF MACHINE LEARNING MODELS FOR PREDICTING THYROID DISEASE



BY

IDRIS BABALOLA

Table of Contents

1. Introduction and Background

Thyroid Disease is a prevalent health condition that affects the proper functioning of the thyroid gland, a vital organ responsible for regulating metabolism and hormone production. Imbalances in thyroid hormones can lead to various health issues, including hyperthyroidism and hypothyroidism (Riajuliislam et al, 2021). According to the World health organization, 200 million people of the world's population are afflicted with enlargement of the thyroid gland (Bayliss & Tunbridge, 1991). Diagnosing thyroid disease accurately is crucial for effective medical interventions and personalized treatment plans. The exploration of health-related datasets has become increasingly crucial in modern healthcare. This work aims to analyze thyroid disease data and unravel valuable insights facilitate early diagnosis and personalized treatment through identification of key feature combinations and development of effective machine learning model for thyroid disease prediction. This approach holds significant societal impact through advancing personalized thyroid disease diagnosis, improving patient well-being and healthcare resources optimization.

Research Question

What role can machine learning play in uncovering the influential features and their combination in developing an effective thyroid disease predictive model?

2. Methods

In providing suitable answer to the research question, exploratory data analysis (EDA) of the provided data was conducted using jupyter notebook and machine learning python packages including pandas, sklearn, scipy and matplotlib. Subsequently, feature selection and data modelling.

i. Data Collection, Cleaning and Exploration

Ten datasets was provided with various dimension ranging from 5-30 features (https://archive.ics.uci.edu/ml/datasets/thyroid+disease). These features include “age”, “sex”, “goitre”, “tumor” as input variables and “negative”, “increased binding protein”, “decreased binding protein” as target variable amongst others. Nine of the dataset are noted to be highly imbalance. For instance, dataset “allbp” contains 2800 observations and the class “negative” alone in the target variable has over 2500 instances. Imbalance data are usually tackled by getting more data, oversampling, undersampling. (Taamneh et al., 2023)

However, only the dataset “new-thyroid.data” presents the best possibility of managing its imbalance. The reason why imbalance was not handled for the other nine datasets and particularly “new-thyroid.data” is because employing techniques such as over/under sampling and gathering more data could potentially lead to overfitting and model not generalizing. (Choirunnisa and Buliali, 2018). This evidence shows the importance of robust data collection methods to minimise errors (Zhong et al, 2021),(Ganesha and Sreeramana Aithal, 2022).
As mentioned, the “new-thyroid.data” dataset with a dimension of 215 instances and 6 attributes (T3-resin, total Serum thyroxin, serum triiodothyronine, thyroid-stimulating hormone (TSH), difference of TSH value and class attribute) was adopted for this work. FIve(5) of these attributes except the class attribute are considered the input variables, while the class attribute consist of three classes namely; “normal”, “hyperthyroidism”, and “hypothyroidism”. Referring back to the data imbalance, the “normal” class consists of 150 instances, while hyperthyroidism and hypothyroidism are 35 and 30 instances respectively.
To avoid garbage in garbage out, appropriate data cleaning is crucial to mitigate noise in our data (Xiong et al, 2006). Column names were assigned to the dataframe to enhance clarity. The presence of any missing values and NaNs was meticulously checked. Fortunately, the dataset

has no missing values and the cleaned dataset was saved as "new-thyroid.csv". In the rare case of encountering NaNs, a KNN imputation method with five nearest neighbors was employed to handle the missing values and ensuring that no valuable information was overlooked.
Additionally, data exploration was conducted. We began by performing correlation analysis to understand the relationships between features, visualizing the correlation. Descriptive statistics were computed to gain an overview of the data distribution. Next, Outliers were observed in most of the features by utilizing box plots, but they were not removed as the models to be experimented, such as Random Forest, are robust to outliers. We further explored the distribution of each feature using histograms, violin plots, and kernel density estimation (KDE) plots. Data mining was further explored by using K-Means clustering approach. Additionally, point biserial correlation coefficients were calculated to determine the relationship between binary and continuous variables, with values greater than or equal to 0.2 indicating significant correlation (Kline, 1986).

ii. Feature Importance and Feature Selection

To identify the most relevant features for predicting the target label, three different methods were employed: Random Forest, Recursive Feature Elimination (RFE), and Lasso Regularization. For each method, we assessed the importance of each feature to the target label. The results from the three techniques were then organized in separate dataframes for easy comparison. The Random Forest method ranked features based on their importance percentage (Cantão et al., 2022), RFE ranked features from most to least important (Rufino et al., 2023), and Lasso Regularization provided feature coefficients (Kumar et al, 2020). By combining the insights from these methods, we can make informed decisions about feature selection, ultimately enhancing the predictive power and interpretability of our model.

iii. Data Transformation

The dataset contains a total of five numerical columns, excluding the 'Target Variable,' which represents the class attribute. We proceed with feature scaling using the StandardScaler to standardize to fit the numerical column between 0 and 1. This step is essential to ensure that all independent features are on the same scale, preventing certain features from dominating others during the modeling process(Ahmed Mia et al., 2021),(Singh Dalwinder & Singh Birmohan, 2019).

iv. Classification Modelling and Evaluation

After data transformation step, the dataset was split into training and testing sets in an 80-20 ratio. Lazy Predict library was initially used to get a quick idea of the model performances, allowing experimentation with 29 different algorithms. The top performing models were identified. Cross-validation technique was used to assess the performance of a top models by dividing the data into five (5) subsets, training and evaluating the model on different combinations, allowing for a more reliable estimation of its generalization ability. The models were optimized using GridSearchCV and the best parameters for each model were identified. GridsearchCV is an exhaustive approach which combines the best parameters to optimize model performance(Uddin et al, 2021),(Belete et al, 2021). The models were retrained and reevaluated with optimized hyperparameters.

3. Results and Discussion of Results

Features	Point Biserial Correlation	P-Values
T3-resin	0.11	0.11
Serum Thyroxin	-0.13	0.06
Serum Triiodothyronine	0.08	0.23
TSH	0.56	0.00
Difference of TSH	0.52	0.00

Table 3.1 Point Biserial Correlation

Based on the Point Biserial Correlation scores presented in Table 3.1, TSH and Difference of TSH stand out as the most important features for the target variable exhibiting notably high correlations (0.56 and 0.52, respectively), indicating a significant relationship with the target. In contrast, T3-resin and Serum triiodothyronine demonstrate relatively weaker correlations (0.11 and 0.08, respectively), while Serum thyroxine exhibits a negative correlation (-0.13) The significance of the features is further assessed by the corresponding P-Values. In our results, TSH and Difference of TSH possess the most substantial predictive power with extremely low P-Values (both at 0.00), confirming their importance in feature selection.

Random Forest		Recursive Feature Elimination		Lasso regularization	
Features	Ranking	Features	Ranking	Features	Ranking
Serum Thyroxine	0.39	Serum Thyroxine	1	TSH	0.05
Difference of TSH	0.20	Serum Triiodothyronine	2	Serum Triiodothyronine	0.05
Serum Triiodothyronine	0.18	TSH	3	Difference of TSH	0.03
TSH	0.18	Difference of TSH	4	Serum Thyroxine	0.01
T3-resin	0.06	T3-resin	5	T3-resin	-0.00

Table 3.2 Feature Selection

The three techniques presented in Table 3.2 provide valuable insights into feature importance for thyroid disease prediction. Serum thyroxine consistently appears as an essential predictor in all three approaches. TSH and Serum triiodothyronine are also consistently ranked high in importance, reinforcing their role in accurate prediction. Despite minor differences in the rankings between the techniques, the consensus on the critical features (Serum thyroxine, TSH, and Serum triiodothyronine) enhances the credibility of these variables for feature selection in thyroid disease prediction models. Similarly, Riajuliislam et al., 2021 performed RFE in several algorithms to see which features are most appropriate for a particular algorithm and found out three important features namely Serum thyroxine, TSH and Serum are consistently rated high in most of the algorithms used.

Evaluation							
without Hyperparameter tuning				**with Hyperparamater tuning**			
Model	Accuracy	Recall	F1 Score	Model	Accuracy	Recall	F1 Score
Gaussian Naïve Bayes	100	100	100	Gaussian Naïve Bayes	100	100	100
Quadratic Discriminant Analysis	100	100	100	Quadratic Discriminant Analysis	97.67	94.44	96.42
Logistic Regression	100	100	100	Logistic Regression	97.67	94.44	96.42

Table 3.3 Model Performance Metrics

	Cross Validation 5Folds						
Model	Data Mode	1	2	3	4	5	Mean Value (%)
Gaussian Naïve Bayes	Train	98	98	97	97	98	97
	Test	95	95	100	100	93	97
Quadratic Discriminant Analysis	Train	97	98	96	96	98	97
	Test	95	93	100	100	93	96
Logistic Regression	Train	97	96	95	97	98	97
	Test	93	98	98	98	88	95

Table 3.4 Cross Validation of Top Three Models

The top three models were selected based on their accuracy, recall and F1 score as presented in Table 3.3. Quadratic Discriminant Analysis, Logistic Regression, and Gaussian Naive Bayes were all identified with accuracy of 100% and F1 score of 100%. Cross-validation with five folds was performed to assess the models' generalization abilities as shown in Table 3.4 which revealed potential signs of overfitting. The Gaussian Naive Bayes model emerged as the best-performing model, with equal mean train accuracy and mean test accuracy of 97% on cross-validation. The hyperparameter of the top three models as shown in table 3.3 significantly improved the cross validated Gaussian Naive Bayes and gave a slight model improvement on the quadratic discriminant analysis and Logistic regression. In a previous work of Salman et al.,

2021, Logistic regression and Gaussian Naive Bayes gave an accuracy of 91.73% and 90.67% respectively. The results show substantial performance in classifications of thyroid disease.

4. Conclusion

In this study, we assessed thyroid disease dataset using twenty-nine (29) different machine learning algorithms. Gaussian Naïve Baye algorithms demonstrated the best prediction performance, achieving 100% accuracy and F1 score. Quadratic Discriminant Analysis and Logistic Regression also performed well, with an optimized F1 score both achieving 96.42%. Serum thyroxine emerged as a crucial feature across all three approaches implemented for feature importance, closely followed by TSH and Serum triiodothyronine, while T3-resin ranked as the least important feature.

These key findings enable researchers to confidently prioritize these features for enhanced prediction accuracy. These findings evidence the potential of machine learning techniques for predicting thyroid disease and analyzing more robust datasets. Although the data analyzed is relatively small, leveraging the blessing of dimensionality could enhance model generalization and enable capturing more complex relationships between features, resulting in accurate predictions with clinical applicability. Further investigations and consideration of domain knowledge can aid in determining the final feature subset for subsequent analysis and modeling.

References

Ahmed Mia, M.Z., Islam, M.M., Haque, M., and Rahman, S.M.A.M. (2021). IRFD: A Feature Engineering based Ensemble Classification for Detecting Electricity Fraud in Traditional Meters. In: Proceedings of the 2021 24th International Conference on Computer and Information Technology (ICCIT). DOI: 10.1109/ICCIT54785.2021.9689842.

Bayliss, R. I. S., & Tunbridge, W. M. G. (1991). Thyroid Disease: The Facts. United Kingdom: Oxford University Press.

Belete, Daniel & D H, Manjaiah. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications. 44. 1-12. 10.1080/1206212X.2021.1974663.

Cantão, A.H., Macedo, A.A., Zhao, L., & Baranauskas, J.A. (2022). Feature Ranking from Random Forest Through Complex Network's Centrality Measures: A Robust Ranking Method Without Using Out-of-Bag Examples. In: Advances in Databases and Information Systems. DOI: 10.1007/978-3-031-15740-0_24.

Choirunnisa, S. and Buliali, J.L. (2018). Hybrid Method of Undersampling and Oversampling for Handling Imbalanced Data. In: Proceedings of the International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) 2018, Yogyakarta. DOI: 10.1109/ISRITI.2018.8864335.

Ganesha, H.R. and Sreeramana Aithal, S. (November 2022). "How to Choose an Appropriate Research Data Collection Method and Method Choice Among Various Research Data Collection Methods and Method Choices During Ph.D. Program in India?" International Journal of Management, Technology, and Social Sciences, 9(2), 233-239. DOI: 10.47992/IJMTS.2581.6012.0233.

Kline, P. (1986). A handbook of test construction: Introduction to psychometric design. Routledge.

Kumar, Agni & Hung, Nancy & Wu, Yuhan & Baek, Robyn & Gupta, Amar. (2020). Predictive Modeling for Telemedicine Service Demand. Telehealth and Medicine Today. 5. 10.30953/tmt.v5.186.

Riajuliislam, Md & Rahim, Khandakar & Mahmud, Antara. (2021). Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques. 60-64. 10.1109/ICICT4SD50815.2021.9397052.

Rufino, J., Ramirez, J.M., Aguilar, J., [...], Anta, A.F. (2023). Feature Selection for an Explainability Analysis in Detection of COVID-19 Active Cases from Facebook User-Based Online Surveys. Preprint. Jun 2023.

Salman, Khalid & Sonuç, Emrullah. (2021). Thyroid Disease Classification Using Machine Learning Algorithms. Journal of Physics: Conference Series. 1963. 012140. 10.1088/1742-6596/1963/1/012140.

Singh, Dalwinder & Singh, Birmohan. (2019). Investigating the impact of data normalization on classification performance. Applied Soft Computing. 105524. 10.1016/j.asoc.2019.105524.

Taamneh, M.M., Taamneh, M.M., Taamneh, S., Taamneh, S., Alomari, A.H., Alomari, A.H., Abuaddous, M., & Abuaddous, M. (2023). Analyzing the Effectiveness of Imbalanced Data

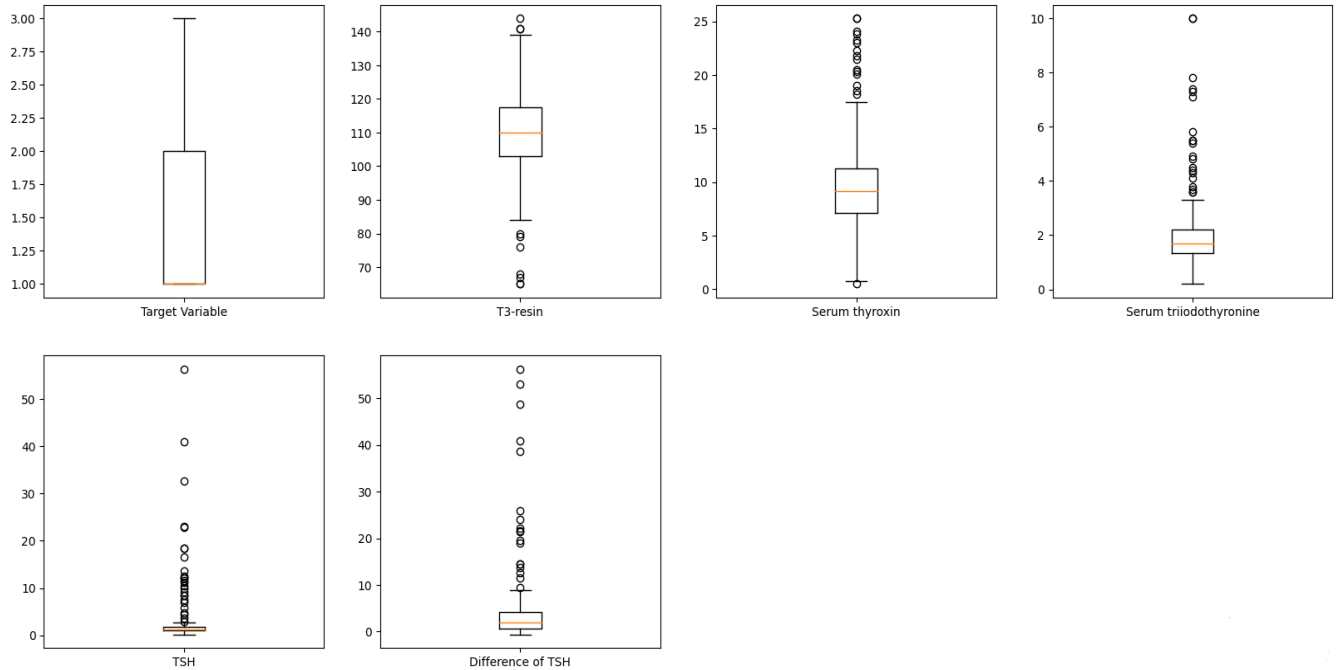
Handling Techniques in Predicting Driver Phone Use. Sustainability, 15(13), 10668. doi: 10.3390/su151310668. License: CC BY 4.0.

Uddin, Md Sihab & Ahmmad Bhuiyan, Erphan & Sarker, Subrata & Das, Sajal. (2021). An Intelligent Short-Circuit Fault Classification Scheme for Power Transmission Line. 10.1109/ACMI53878.2021.9528200.

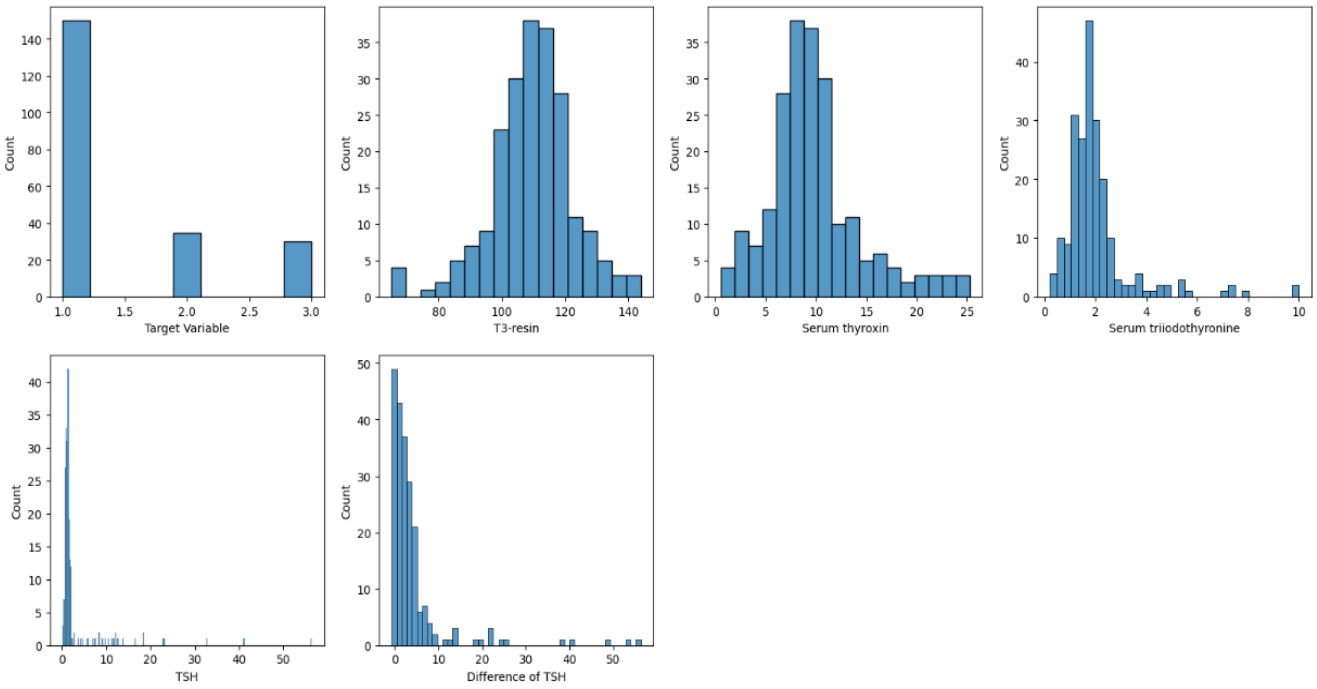
Xiong, Hui & Pandey, Gaurav & Steinbach, Michael & Kumar, Vipin. (2006). Enhancing data analysis with noise removal. Knowledge and Data Engineering, IEEE Transactions on. 18. 304-319. 10.1109/TKDE.2006.46.

Zhong, Rujie & Hanna, Josiah & Schäfer, Lukas & Albrecht, Stefano. (2021). Robust On-Policy Data Collection for Data-Efficient Policy Evaluation.

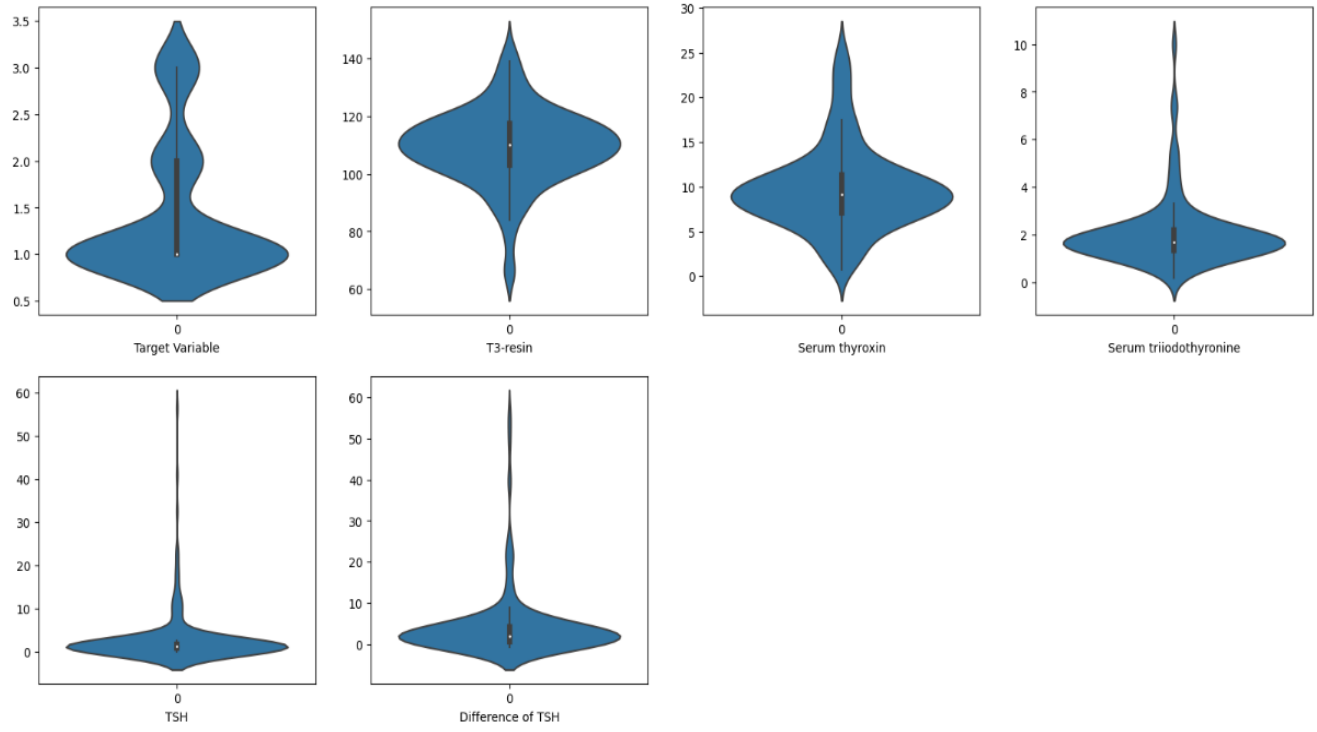
Appendix



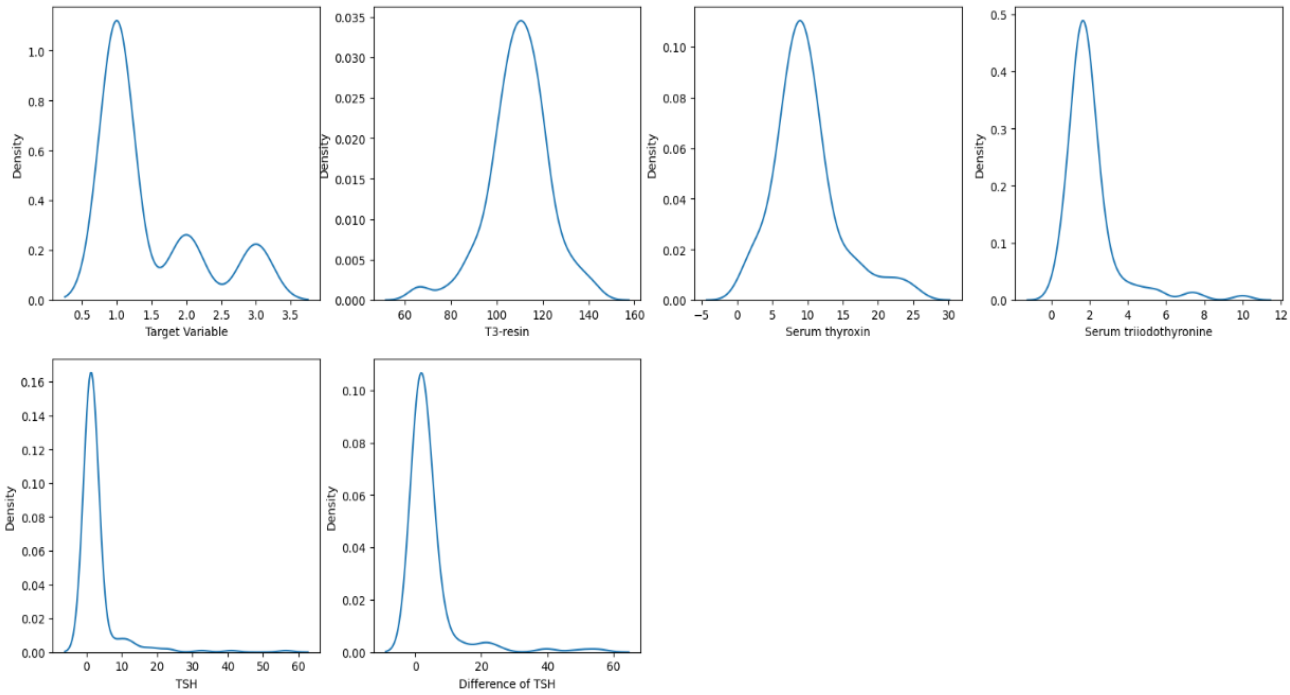
Outlier plot



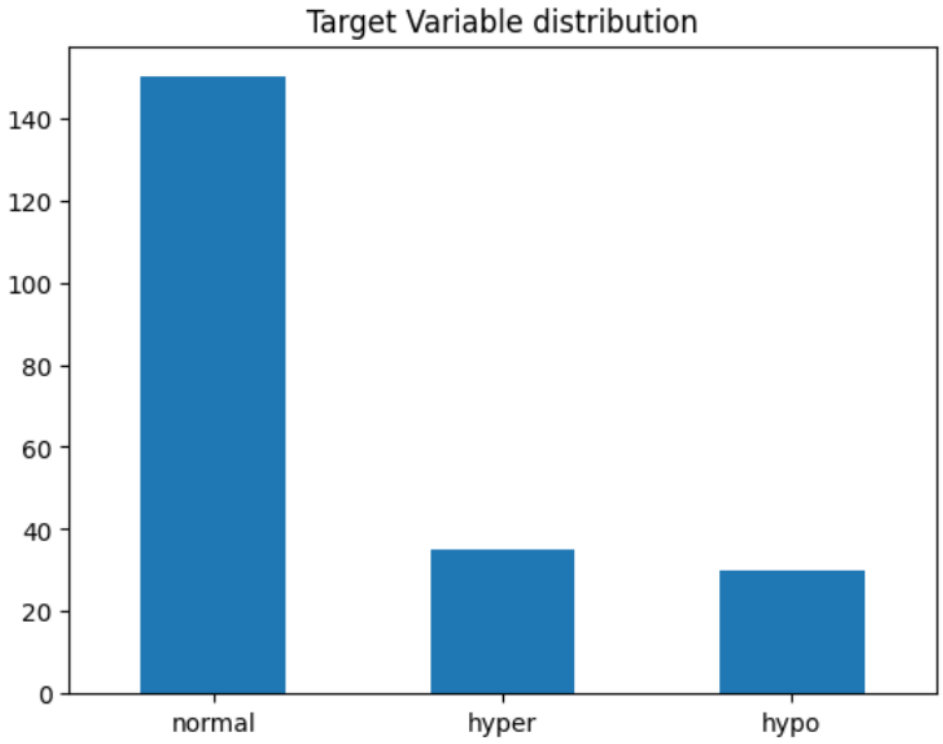
Histogram Plot



Violin Plot



Kernel Density Estimate(KDE) Plot



Target variable class balance