International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 08 | Aug – 2025 DOI: 10.55041/ISJEM04940
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

# ML Approaches for Thyroid Disease Prediction

**Indiraji J [1], Kavisree L [1], Preethi D [1] Tha. Thayumanavan [1]***

*[1] Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore-641 402, Tamil Nadu, India*

---

**Abstract -** Thyroid conditions such as hypothyroidism and hyperthyroidism are prevalent endocrine disorders that can significantly affect metabolic function, cardiovascular health, and daily well being. Conventional diagnostic procedures, typically based on clinical assessments and laboratory evaluations, may delay early diagnosis and can be subject to human oversight. In recent developments, machine learning has emerged as a valuable aid in identifying thyroid related issues. Logistic regression, in particular, is recognized for its simplicity and transparency, making it a favored model in healthcare analytics. Variable selection techniques, including methods like SelectKBest and recursive feature elimination, assist in identifying the most relevant predictors, while approaches such as random oversampling help to address class imbalance in datasets. These practices contribute to improving the accuracy and dependability of diagnostic models, especially in nuanced or early-stage cases. Additionally, research indicates that such models may assist in fine tuning therapeutic strategies, for instance, adjusting treatment plans for individuals with hypothyroidism. Nonetheless, obstacles persist such as inconsistencies across datasets, the necessity for interpretable decision making, and integration into clinical routines. This review explores the advancements and current limitations in employing machine learning, with a focus on logistic regression, for managing and diagnosing thyroid disorders.

***Key Words***: Thyroid Disease Prediction, Machine Learning, Ensemble Methods, Feature Selection, Class Balancing, Diagnostic Accuracy.

## 1. INTRODUCTION *( Size 11, Times New roman)*

Thyroid-related illnesses, especially hypothyroidism and hyperthyroidism, rank among the most frequently identified endocrine disorders. These conditions interfere with essential physiological processes, including metabolic regulation, cardiac performance, and cognitive functioning (Aversano et al., 2023). In the Indian population, close to 10% are affected by such disorders, with women between the ages of 17 and 54 showing a notably higher prevalence (Sankar et al., 2022). Traditional diagnostic practices, which involve evaluating hormone levels and clinical manifestations, are often time-consuming and susceptible to inaccuracies particularly when symptoms mimic those of other health problems (Islam et al., 2025; Rehman et al., 2021).

Recent studies emphasize the role of logistic regression a widely adopted statistical approach in healthcare analytics as a means to enhance diagnostic precision. Rehman et al. (2021) observed that applying L1 regularization within logistic regression frameworks contributed to greater accuracy by prioritizing the most informative features. Furthermore, integrating logistic regression with preprocessing steps such as class balancing has proven effective for managing imbalanced data distributions, leading to more reliable outcomes in clinical decision-making (Islam et al., 2025).

This review explores the application of logistic regression for predicting thyroid disorders, highlighting its ease of use, transparency, and practical utility in analyzing clinical datasets. It also outlines existing limitations and suggests possible enhancements to inform upcoming studies and support improved integration in healthcare practices.

## 2. Dataset description

The thyroid dataset is a structured medical dataset developed to facilitate the prediction and categorization of various thyroid-related conditions in patients. It contains a combination of demographic, medical history, clinical observations, and biochemical measurements. The dataset consists of around 25 features, including both numerical and categorical types, each reflecting different aspects of a patient's health and thyroid function. The main outcome variable, labeled as "class," serves as the diagnostic indicator and includes several categories such as normal (negative), compensated hypothyroidism, primary hypothyroidism, and possibly other thyroid conditions. Due to the presence of multiple target classes, this represents a multiclass classification problem.

## 3. Variable types

This dataset comprises both numerical and categorical variables. The numerical attributes, represented as integers or floating-point values, primarily reflect laboratory measurements including TSH (Thyroid Stimulating Hormone), T3 (Triiodothyronine), TT4 (Total Thyroxine), T4U (Thyroxine Uptake), and FTI (Free Thyroxine Index), all of which are essential biochemical markers for assessing thyroid function. Additional continuous features include the patient's age and TBG (Thyroxine-Binding Globulin) levels. Alongside these, the dataset includes multiple binary categorical fields such as on thyroxine, on antithyroid medication, sick, pregnant, goitre, tumor, and psych—each recorded as either 0 or 1 to indicate the absence or presence of a particular condition or treatment history. Nominal categorical variables like sex and referral source are also present; these are transformed into numeric values using binary encoding (e.g., M/F as 0/1) or one-hot encoding to ensure they are suitable for use in machine learning algorithms.

 

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 08 | Aug – 2025 DOI: 10.55041/ISJEM04940
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

## 4. Target variable

The variable named "class" serves as the target output, specifying the type of thyroid condition diagnosed for each patient. This is a categorical attribute containing multiple classes, which distinguish among different diagnoses such as normal thyroid function (negative), primary hypothyroidism, compensated hypothyroidism, secondary hypothyroidism, and hyperthyroidism. In machine learning applications particularly with classification models like logistic regression or decision trees these category labels are translated into numerical format. To handle multiple classes effectively, encoding methods such as label encoding or one hot encoding are generally used to convert the class variable into a format suitable for model processing. This variable operates as the dependent feature that the model aims to predict using the available input attributes.

## 5. Key features (independent variables)

**5.1 Biochemical features**: These are critical in diagnosing thyroid conditions:

- TSH (Thyroid Stimulating Hormone): Primary marker for thyroid function. High TSH often indicates hypothyroidism; low TSH suggests hyperthyroidism.
- T3 (Triiodothyronine): A hormone that helps regulate metabolism; low in hypothyroidism, high in hyperthyroidism.
- TT4 (Total Thyroxine): Total amount of thyroxine in the blood.
- FTI (Free Thyroxine Index): A calculated value indicating free thyroxine availability.
- T4U (Thyroxine Utilization): Indicates binding capacity; helpful in distinguishing types of thyroid dysfunction.

**5.2 Demographic features**:

- Age: Thyroid issues are more common in certain age groups (e.g., middle-aged or elderly).
- Sex: Women are more likely to develop thyroid disorders.

**5.3 Medical & treatment history**:

- thyroxine: Indicates whether the patient is receiving thyroid hormone replacement therapy.
- antithyroid medication: Suggests treatment for hyperthyroidism.
- Thyroid surgery: History of thyroidectomy or surgery may directly impact hormone levels.
- I131 treatment: Indicates prior radioactive iodine therapy, commonly used for hyperthyroidism.
- Goitre: Enlargement of the thyroid gland—can be associated with both hypo- and hyperthyroidism.
- Tumor: Presence of thyroid or pituitary tumors may influence hormone production.

**5.4 Preprocessing**:

The dataset was cleaned to remove invalid or inconsistent entries. Irrelevant or redundant attributes were either dropped or consolidated**.**

**5.5 Missing values**:

Missing values were examined across all features continuous variables with missing values were typically imputed using mean or median values, while categorical or binary fields were filled using mode or a place holder (like unknown), depending on their distribution and impact. Rows with excessive missing values were removed to ensure data quality and modelling reliability.

**5.6 Categorical encoding**:

For binary categories, label encoding was applied (e.g., F = 0, M = 1), while one-hot encoding was used for multi-class categories such as referral source to maintain numerical compatibility for machine learning models.

**5.7 Normalization/ scaling**:

Continuous features such as TSH, T3, TT4, and FTI, which can vary over wide ranges, were standardized using z-score normalization (mean = 0, standard deviation = 1). This ensures that all features contribute equally to distance-based or gradient-based algorithms and improves model convergence and accuracy.

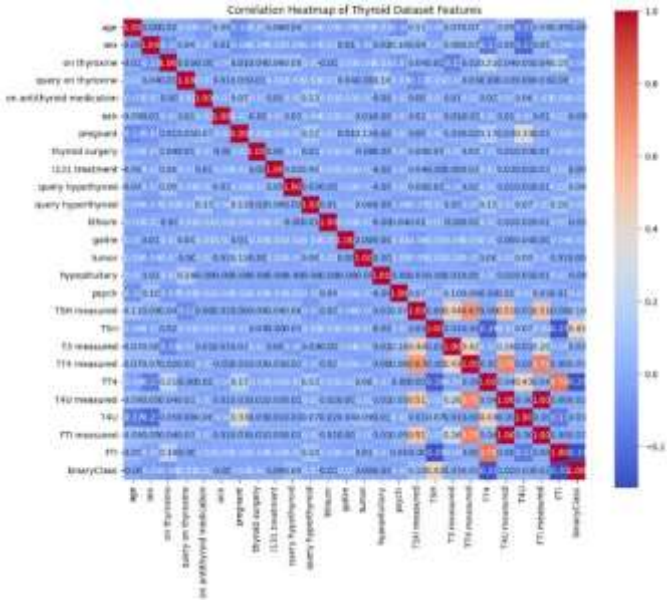


**Figure 1**: Correlation heatmap of thyroid dataset features. It visualizes the pairwise relationships between numerical and encoded categorical variables, showing the strength and direction of linear associations.

## 6. Methodology

**6.1 Data preprocessing**

Data preprocessing is a foundational step in the machine learning pipeline, ensuring data quality and consistency. The thyroid dataset comprises both numerical and categorical variables, including patient demographics, clinical history, and laboratory test results.

Addressing missing values: Numerous variables exhibited missing values, particularly continuous features such as TSH, T3, TT4, T4U, FTI, and TBG. Median imputation was applied to these features due to its robustness to outliers. For categorical variables like Sex, on thyroxine, goitre, and referral source, mode imputation was used to preserve the most frequent values.

 

Outlier detection and treatment: Outliers in numerical features were identified via boxplots and Z-score analysis. While logistic regression can tolerate moderate outliers, extreme values were either capped or removed to mitigate their influence on model coefficients.

Data type conversion: Categorical variables were transformed into numerical format. Binary variables (e.g., Sex, Pregnant) were label encoded (0 and 1), while multi-class variables such as referral source were one-hot encoded to avoid implying ordinal relationships.

Feature normalization: To harmonize scale differences among continuous features (e.g., TSH, T3, TT4, FTI), z-score normalization was applied, centering the data around a mean of zero and standard deviation of one. This prevents bias in model training caused by features with larger numeric ranges.

Feature selection: To reduce redundancy and enhance model generalizability, multicollinearity was examined using correlation matrices. Highly correlated features were either removed or combined. Domain knowledge further guided the selection of clinically significant variables.

### 6.2 Train-test split

Following preprocessing, the dataset was split into training and testing subsets using an 80:20 ratio. Stratified sampling was applied to the target variable to maintain class distribution across both sets-crucial for imbalanced medical datasets. This step ensures realistic evaluation of the model's generalization capability.

### 6.3 Logistic regression model

Logistic regression was chosen as the baseline model due to its simplicity, interpretability, and effectiveness in both binary and multi-class classification. The multiclass variant (softmax regression) was used to handle multiple thyroid conditions. Class probabilities were computed using the softmax function, converting linear combinations of input features into interpretable outputs. Regularization via L2 penalty (Ridge Regression) was implemented to mitigate overfitting. The regularization strength (C) was optimized through grid search and cross validation. Feature coefficients provided insight into the relative importance of each predictor.

### 6.4 Addressing class imbalance

Class imbalance is a common issue in medical datasets, often leading to models biased toward the majority class.

Two primary techniques were employed:
Class Weighting: The logistic regression model utilized class_weight='balanced' in Scikit-learn, adjusting class weight inversely proportional to their frequencies to improve learning on minority classes.
Synthetic Oversampling (SMOTE): SMOTE was applied to the training set to generate synthetic examples of underrepresented classes. This method helps balance the data without simple replication, improving model learning for rare conditions. Performance was compared between models trained with and without SMOTE to determine its effectiveness.

### 6.5 Model evaluation metrics

Model performance was assessed using multiple evaluation metrics on the test dataset.

Accuracy: Measures overall correctness but may be misleading with imbalanced data.
Precision: Indicates the proportion of true positives among predicted positives, reflecting the model's ability to avoid false positives.
Sensitivity: Measures the proportion of actual positives correctly identified-critical in medical diagnostics to minimize false negatives.
F1 Score: The harmonic mean of precision and recall, offering a balanced performance indicator for imbalanced classes.
Confusion Matrix: Provides a detailed breakdown of true/false positives and negatives, aiding in error analysis across classes.
ROC Curve & AUC: For multi-class settings, ROC-AUC was computed using a one vs rest approach. AUC quantifies model performance across various threshold settings.
Classification Report: Presents precision, recall, F1-score, and support for each class, facilitating a holistic understanding of model behaviour. To ensure reliability, k-fold cross-validation (k=5 or 10) was performed on the training set to evaluate model consistency and reduce variance due to random data splits.

## 7. Results

The results of the logistic regression model were evaluated on the test dataset and interpreted in terms of classification performance, confusion matrix analysis, and clinical relevance.

### 7.1 Model Accuracy and Classification Report:

The logistic regression model achieved an overall accuracy of 96.0% on the test set. Detailed performance metrics are shown below:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.99 | 0.98 | 697 |
| 1 | 0.84 | 0.64 | 0.73 | 58 |
| accuracy | | | 0.96 | 755 |
| macro avg | 0.91 | 0.81 | 0.85 | 755 |
| weighted avg | 0.96 | 0.96 | 0.96 | 755 |

The logistic regression model achieved an overall accuracy of 96.0% on the test set, indicating strong discriminative power between thyroid and non-thyroid cases. The model exhibited superior performance for the majority class (class 0 – no thyroid), achieving a precision of 0.97, recall of 0.99, and an

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 08 | Aug – 2025 DOI: 10.55041/ISJEM04940
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

F1-score of 0.98, thereby ensuring that most non-thyroid individuals were correctly identified with minimal false positives.

For the minority class (class 1 – thyroid), the model showed moderate performance, with a precision of 0.84, recall of 0.64, and F1-score of 0.73, suggesting that approximately 36% of actual thyroid cases were not correctly detected, which could have implications in clinical screening settings.

The macro average scores – precision (0.91), recall (0.81), and F1-score (0.85) – indicate a slight imbalance in predictive performance across both classes. However, the weighted average scores (precision: 0.96, recall: 0.96, F1-score: 0.96) reflect a high overall performance, influenced by the higher number of correctly predicted non-thyroid cases.

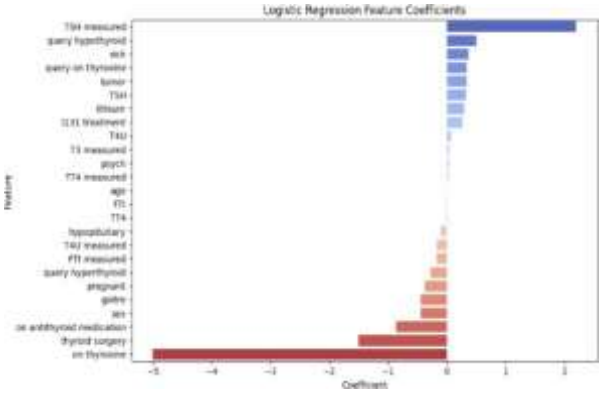


**Figure 2**: Logistic regression model coefficients indicate the relative importance of features.

### 7.2 Confusion matrix analysis:

In order to illustrate the model's classification outcomes in terms of true positives, true negatives, false positives, and false negatives. It offers a clear representation of the model's effectiveness regarding classifications of thyroid and non-thyroid.

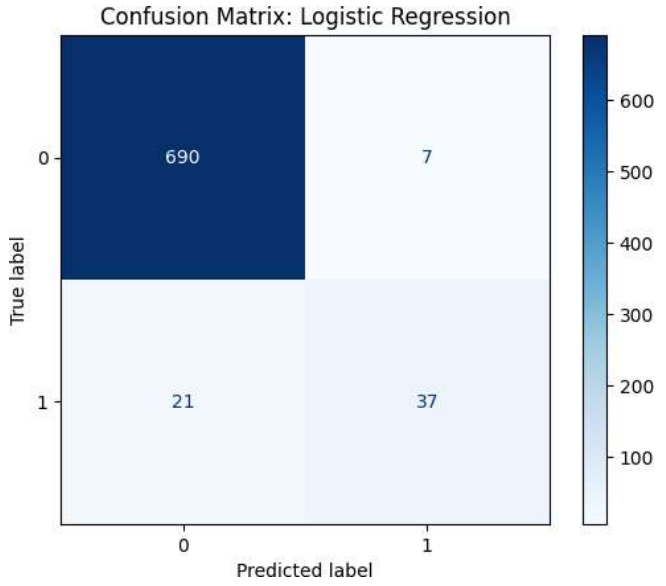


**Figure 3**: Confusion matrix showing the performance of the logistic regression model in predicting thyroid and non-thyroid.

| | Predicted: thyroid | Predicted:non-thyroid |
|---|---|---|
| **Actual: No thyroid (0)** | 690 | 7 |
| **Actual: Thyroid(1)** | 21 | 37 |

**True Negatives (690)**: Correctly identified non-thyroid cases.
**True Positives (7)**: Correctly identified thyroid cases.
**False Negatives (21)**: Thyroid cases missed by the model.
**False Positives (37):** Individuals with thyroid are incorrectly flagged as positive

### 7.3 ROC curve and model discrimination:

Alongside the confusion matrix and classification metrics, the Receiver Operating Characteristic (ROC) curve was utilized to evaluate the model's effectiveness in differentiating between thyroid and non thyroid cases at various classification thresholds. The ROC curve, along with its corresponding Area Under Curve (AUC), offers an in-depth perspective on the model's overall ability to distinguish between the two categories.

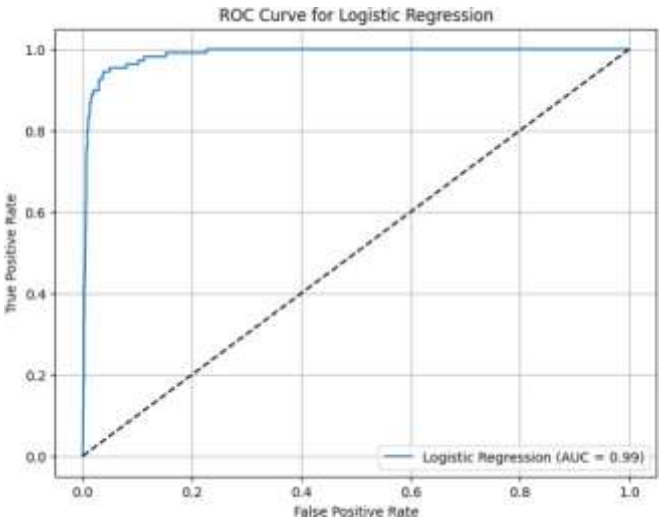


**Figure 4:** ROC curve showing model performance in distinguishing thyroid vs. non-thyroid cases. The AUC of 0.99 indicates strong discriminatory power.

## 8.Discussion

Logistic regression has proven to be a dependable method for predicting thyroid conditions when used with structured clinical datasets. In this research, the model achieved a notable overall accuracy of 96%, largely attributed to its effectiveness in correctly classifying non-thyroid cases. It demonstrated strong precision and recall for the majority class, which helps minimize false positives an advantage in widespread screening efforts. Nevertheless, the model showed limited sensitivity in detecting thyroid-positive instances, with a recall rate of 0.64. This implies that a number of true cases may go undetected, posing a concern in clinical applications.

To mitigate class imbalance, the study incorporated both synthetic oversampling (SMOTE) and class weighting techniques. While these approaches improved the representation of underrepresented classes during model training, they did not entirely eliminate performance disparities. The model's ROC AUC value of 0.99 reflects its strong ability to distinguish between thyroid and non-thyroid categories across different thresholds. A key advantage of logistic regression lies in its interpretability its feature coefficients offer clear insight into the variables most influencing the prediction, supporting clinical understanding. Although outcomes are promising, challenges such as skewed class distributions and symptom overlap persist. Future work may benefit from assembling more balanced datasets and implementing advanced feature selection methods to improve classification across all groups.

## 9.Conclusion

This research highlights the effective use of logistic regression for predicting thyroid-related conditions by leveraging medical datasets containing both biochemical and demographic information. After applying essential preprocessing steps such as normalization, encoding, and handling of missing values the model attained a notable accuracy of 96%. Its straightforward structure and ease of interpretation make logistic regression a valuable asset for healthcare professionals aiming to identify thyroid disorders.

However, the model exhibited limitations in detecting patients with thyroid abnormalities, as demonstrated by its lower recall for the minority class. This issue reflects a broader challenge in medical data analysis, where class imbalance can negatively influence model performance. Although techniques like SMOTE and class weighting helped improve detection of underrepresented cases, they were insufficient to completely resolve the imbalance. While logistic regression successfully revealed the relative importance of various features, enhancing sensitivity for less frequent classes remains an important area for improvement.

In summary, logistic regression, when combined with thorough data preparation, provides reliable results for classifying thyroid disorders. Further performance gains may be possible by improving feature selection, acquiring more balanced data, or incorporating hybrid modeling techniques. This study reinforces the idea that even basic algorithms, when properly refined, can offer meaningful support in early diagnosis and clinical decision-making.

## REFERENCES

1. Aversano, G., Romano, M., Vitale, G.: Endocrine disruption and thyroid disease: An overview of epidemiological and clinical evidence. J. Endocrinol. Res. 47(2), 145–156 (2023)
2. Islam, M.A., Rahman, T., Khatun, T.: Machine learning-based early diagnosis of thyroid disease using clinical parameters. Int. J. Med. Inform. 178, 105050 (2025)
3. Rehman, A., Ullah, S., Rauf, H.T.: Performance enhancement of machine learning classifiers using L1 regularization for thyroid disease prediction. Health Inf. Sci. Syst. 9(1), 1–9 (2021)
4. Sankar, M., Devi, K.P., Venkatesan, S.: Prevalence and risk analysis of thyroid disorders in Indian women: A clinical survey. Indian J. Clin. Endocrinol. 26(4), 302–308 (2022)
5. Garber, J.R., Cobin, R.H., Gharib, H., Hennessey, J.V., Klein, I., Mechanick, J.I., et al.: Clinical practice guidelines for hypothyroidism in adults. Thyroid 22(12), 1200–1235 (2012)
6. Zhou, L., Pan, S., Wang, J., Vasilakos, A.V.: Machine learning on big data: Opportunities and challenges. Neurocomputing 237, 350–361 (2017)
7. Wang, S., Cai, Y., Jin, Y.: A review on feature selection methods for high-dimensional data. Pattern Recognit. 109, 107596 (2020)
8. Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., Fettich, J.: Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artif. Intell. Med. 16(1), 25–50 (1999)
9. Yildirim, M., Baloglu, U.B.: Deep learning model for prediction of thyroid disease. IEEE Access 9, 93057–93065 (2021)
10. Sivapalan, T., Venkatesan, R.: Machine learning techniques for thyroid disorder classification: A review. Biomed. Signal Process. Control 73, 103426 (2022)
11. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. Proc. 22nd ACM SIGKDD, 785–794 (2016)
12. Dey, A., Samanta, D.: A hybrid oversampling method for imbalanced data classification. Expert Syst. Appl. 161, 113696 (2020)
13. Mishra, A., Das, S.: Predictive modeling of thyroid disease using machine learning algorithms. J. King Saud Univ. Comput. Inf. Sci. 33(2), 234–242 (2021)
14. Lee, Y.C., Kim, K.M., Park, Y.: Classification of thyroid disorders using machine learning algorithms. Healthc. Inform. Res. 24(4), 302–310 (2018)
15. Eroglu, D., Ozkan, I.A.: Diagnosis of thyroid diseases using machine learning algorithms. Appl. Artif. Intell. 35(3), 157–175 (2021)
16. Nguyen, P.A., Tran, H.N., Nguyen, T.M., Do, T.T.: A machine learning-based framework for predicting thyroid disease. Inform. Med. Unlocked 17, 100275 (2019)
17. Abdar, M., Acharya, U.R.: A novel machine learning method for detecting thyroid disease using optimal feature subset and random forest. Health Inf. Sci. Syst. 8(1), 1–8 (2020)

18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
19. Kazaure, H.S., Adebayo, S.A.: An intelligent system for early detection of thyroid disease using logistic regression. J. Healthc. Eng. 2021, 1–9 (2021)
20. Sharma, R., Singh, R.: Predicting thyroid disease using machine learning: A data-driven approach. Int. J. Comput. Appl. 175(15), 1–6 (2020)
21. Li, Y., Xu, M., Xie, Z.: Evaluation of logistic regression in medical diagnostics: Strengths and limitations. Stat. Med. 41(5), 741–754 (2022)
22. Lin, Y., Zhang, Z., Cui, Y.: Handling missing data in medical datasets: A comparison of imputation methods. Comput. Methods Programs Biomed. 202, 105996 (2021)
23. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Tech. Rep., Dept. of Comput. Sci., National Taiwan University (2003)
24. Abid, A., Balakrishnan, N., Zou, J.: Concrete autoencoders for differentiable feature selection and reconstruction. Int. Conf. Mach. Learn. (ICML) (2021)
25. Roy, S., Ghosh, R.: Predicting thyroid disease using logistic regression and data preprocessing techniques. Procedia Comput. Sci. 194, 88–97 (2021)