iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61176

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Machine Learning Techniques for the Classification of Thyroid Disease

Parashuram S. Vadar[1], Urmila R. Pol[2],Tejashree T. Moharekar[3]
Department of Computer Science, YCSRD, Shivaji University, Kolhapur,Shivaji University, Kolhapur, Shivaji University, Kolhapur

Abstract: ***In the healthcare system, dealing with a large amount of data is challenging. The techniques of machine learning are used in dealing with such data. As per NFHS – 5 statistics, thyroid diseases are increasing in India. Roughly 1 in 10 Indian adults suffer from a thyroid disorder. It has been expected that more than 42 million peoples suffer from thyroid disease. For the proper diagnosis of disease, it is vital to process the medical data accurately. This study classified thyroid disease cases into hyperthyroid, euthyroid, hypothyroid, and sick. This paper aims to inspect Logistic regression for multiclass categorizing the thyroid dataset. This logistic regression model is evaluated based on its precision, recall, F measure, ROC, RMS Error, and accuracy metrics. Based on the thyroid dataset, we find that logistic regression using the One-vs.-Rest heuristic is 85% accurate, while logistic regression using the multinomial is 86% accurate.***

Keywords: Thyroid, Machine Learning, Logistic Regression, Hyperthyroidism, Hypothyroidism, Sick, Multiclass Classification, Prediction.

I. INTRODUCTION

An imbalance in thyroid hormones can result in thyroid disorders. The thyroid releases Triiodothyronine (T3) and thyroxin (T4) into the bloodstream. The function of the thyroid gland is to make hormones essential for the body's metabolism, growth, and development.

Hyperthyroidism and Hypothyroidism are the two main forms of thyroid disease. Due to the overactive condition of the thyroid gland, hyperthyroidism causes an excessive amount of thyroid hormone to be released into the bloodstream. Symptoms of hyperthyroidism contain unexpected weight loss, rapid or irregular heartbeats, sweating, and irritability.

A condition in which the thyroid gland doesn't produce sufficient thyroid hormone is known as Hypothyroidism. In Hypothyroidism, Heart rate, body temperature, and all aspects of metabolism are disrupted due to the deficiency of thyroid hormones. Symptoms of Hypothyroidism depend on how severe a deficiency in thyroid hormone production is. Symptoms of Hypothyroidism varies person to person. Most people will have some combination of the symptoms like Fatigue, Weakness, Weight gain or increased difficulty losing weight, Coarse, dry hair, Dry, rough pale skin, Hair loss, Cold intolerance, Muscle cramps, Constipation, Depression, Memory loss, etc.

An important challenge in the diagnosis of thyroid disease disorders is the correct interpretation and clinical analysis of thyroid disease disorder datasets. Thyroid prediction methodologies will allow us to reduce the number of characteristics used to classify thyroid disorders.

II. REVIEW OF LITERATURE

Several classification algorithms used in machine learning are compared in this paper along with an Artificial Neural Network (ANN). A number of algorithms were used for classification, including Naive Bayes, Support Vector Machines, k-Nearest Neighbors, Random Forest Classifiers, Logistic Regressions, and also Artificial Neural Networks. An observation was made that the ANN has the highest accuracy of 96.7573%, whereas the Logistic Regression has a good accuracy of 96.1929%, and the ANN will become more computationally expensive if the dataset is larger [5].

Using machine learning techniques, this study aims to develop a classification model for assessing euthyroidism, hyperthyroidism, and hypothyroidism in males, females, and children. The classification of real data is performed using a variety of machine learning algorithms, such as nave bayes, decision trees, random forests, and logistic regression. Metrics such as precision, recall, specificity, and sensitivity have been used to evaluate the accuracy of each of the techniques. The proposed model was trained on thyroid data collected from two hospitals in Haryana between January 2020 and July 2020. In comparison to all four traditional algorithms, the proposed algorithm has an accuracy of 94% [9].

©IJRASET: All Rights are Reserved

The purpose of this study is to develop a predictive model for thyroid diseases using three machine learning classification algorithms, namely K-Nearest Neighbor (KNN), Naive Bayes, and Decision Trees. A dataset from UCI's machine learning repository is used for the experiments. Through 10-fold cross-validation, the performance of the three algorithms is evaluated on several parameters, including Accuracy, Precision, F-Measure, and Recall. Compared to Nave Bayes and KNN, the decision tree offered the highest accuracy rate, with 99.7% [10].

Based on parameters established from the dataset, this paper compares various machine learning algorithms such as decision trees, random forests, KNNs, and artificial neural networks. For accurate classification, the dataset has been manipulated. For better comparison of the datasets, both sampled and unsampled datasets were classified. After manipulating the dataset, the authors obtained the highest accuracy for the random forest algorithm, which was 94.8% accuracy and 91% specificity [8].

This study demonstrated the intuitive understanding of predicting thyroid disease, along with applying logistic regression, decision trees, and kNN as tools for classifying thyroid disorders. In order to accomplish this, thyroid data from the machine learning repository has been retrieved from UC Irvin knowledge discovery in databases. The accuracy of Logistic regression, Decision tree and k-NN classifier is 81.25, 87.5, and 96.875, respectively [1].

In this study, an extensive analysis of different classifiers is presented, including K-nearest neighbor (KNN), Naive Bayes, support vector machines, decision trees and logistic regression implementations with and without feature selection. Due to the additional features of pulse rate, body mass index, and blood pressure, the thyroid dataset was unique and different from other existing studies. There were three iterations in the experiment; the first iteration did not employ feature selection, while the second and third used L1-, L2-based feature selection techniques. A number of factors were evaluated and analyzed, including accuracy, precision, receiver operating curve, and area under curve. Based on the results, classifiers using L1-based feature selection were significantly more accurate (Naive Bayes 100%, logistic regression 100%, and KNN 97.84%) than those without feature selection and using L2-based feature selection [7].

This study uses classification Predictive Modelling followed by binary classification using Decision Tree ID3 and Naive Bayes Algorithms to predict thyroid disease. Using the Decision Tree algorithm, the presence of thyroid in the patient is determined from the Thyroid Patient dataset. If thyroid is present, the Nave Bayes algorithm is applied to determine thyroid stage [2].

This disease is diagnosed through thyroid blood tests, which are usually blurred and noisy. The data was cleansed so the analytics could show the risk of patients getting this illness. ANN Artificial Neural Networks, KNN K-nearest neighbours, decision trees, logistic regressions, and SVM support vector machines are used to predict thyroid disease risk [4].

LT4 treatment trends for hypothyroid patients are predicted in this study. Datasets were created including information on patients treated at the Naples "AOU Federico II" hospital. Various machine learning algorithms were used in this study. Ten different classifiers were compared. There are good results from the different algorithms, especially with the Extra-Tree Classifier, where accuracy reaches 84% [3].

Early detection of thyroid disease is crucial to preventing fatal thyroid diseases like thyroid cancer. Machine learning (ML) has become a reliable component for thyroid disease prediction. The model classifier was trained and tested using datasets from the University of California, Irvine (UCI). The dataset was used to implement several machine learning algorithms and their confusion matrices were presented. After a detailed comparison of accuracy, precision, sensitivity, F1 score, ROC-AUC, it was conclusively determined that Multilayer Perceptron (MLPC) had the highest accuracy of 99.70% after hyperparameter optimization [6].

III. METHODOLOGY

The dataset for this classification was taken from the UCI repository. It consists of 3221 records and 27 attributes. Data pre-processing was performed to eliminate missing values from the dataset, and attribute type casting was done whenever necessary. As shown below, Category is a nominal variable in the Thyroid dataset with four different values.

```
['negative', 'hyperthyroid', 'hypothyroid', 'sick']
```

The logistic regression algorithm is a machine-learning technique that allocates records into discrete classes. Currently, this classifier is being used in the study. As a result of linear regression, continuous number values are produced as an output. In logistic regression, predictions are discrete (only exact values or categories are permitted). For binary logistic regression, this will take two values, 0 or 1.

$$Y=b0+b1X+e$$

A sigmoid function is used to map predicted values to probabilities. A real value can be mapped into a value between 0 and 1 using this function. In machine learning, sigmoid functions are used to map predictions to probabilities.

$$\sigma t = 11 + e - t$$

Based on X being given as input, the estimated probability tells how confident we can be that the predicted value will be the actual value.

Following is a description of the descriptive statistics for the thyroid dataset.

	count	mean	std	min	25%	50%	75%	max
Age	3221.0	52.406085	19.104151	1.000	37.00	55.00	68.00	94.00
FTI	3221.0	0.306116	0.460950	0.000	0.00	0.00	1.00	1.00
FTI Measured	3221.0	0.106489	0.308510	0.000	0.00	0.00	0.00	1.00
Goitre	3221.0	0.014902	0.121180	0.000	0.00	0.00	0.00	1.00
Hypopituitary	3221.0	0.010866	0.103689	0.000	0.00	0.00	0.00	1.00
I131 Treatment	3221.0	0.043775	0.204626	0.000	0.00	0.00	0.00	1.00
Lithium	3221.0	0.014281	0.118666	0.000	0.00	0.00	0.00	1.00
On Antithyroid Medication	3221.0	0.012729	0.112120	0.000	0.00	0.00	0.00	1.00
On Thyroxine	3221.0	0.017386	0.130725	0.000	0.00	0.00	0.00	1.00
Pregnant	3221.0	0.065508	0.247458	0.000	0.00	0.00	0.00	1.00
Psych	3221.0	0.063955	0.244711	0.000	0.00	0.00	0.00	1.00
Query Hyperthyroid	3221.0	0.004967	0.070315	0.000	0.00	0.00	0.00	1.00
Query Hypothyroid	3221.0	0.008382	0.091186	0.000	0.00	0.00	0.00	1.00
Query on Thyroxine	3221.0	0.028252	0.165718	0.000	0.00	0.00	0.00	1.00
Sex	3221.0	0.000621	0.024915	0.000	0.00	0.00	0.00	1.00
Sick	3221.0	0.045948	0.209406	0.000	0.00	0.00	0.00	1.00
T3	3221.0	0.923316	0.266131	0.000	1.00	1.00	1.00	1.00
T3 Measured	3221.0	6.322330	26.543102	0.005	0.58	1.50	3.00	478.00
T4U	3221.0	0.817138	0.386614	0.000	1.00	1.00	1.00	1.00
T4U Measured	3221.0	1.951770	0.839899	0.050	1.60	1.90	2.20	10.60
TSH	3221.0	0.955914	0.205317	0.000	1.00	1.00	1.00	1.00
TSH Measured	3221.0	107.551630	38.091518	2.000	86.00	102.00	123.00	430.00
TT4	3221.0	0.914312	0.279946	0.000	1.00	1.00	1.00	1.00
TT4 Measured	3221.0	0.988229	0.185982	0.310	0.88	0.97	1.07	2.12
Thyroid Surgery	3221.0	0.914933	0.279024	0.000	1.00	1.00	1.00	1.00
Tumor	3221.0	110.261347	35.967317	2.000	93.00	106.00	123.00	395.00

Fig. 1. Descriptive statistics for the thyroid dataset

In the figure above, we can see that T3 Measured, TSH Measured, and Tumor are all numerical values, whereas the remaining attributes are nominal. Each can only have two possible values. We can also observe that the average age of the patients was 52.4, indicating that most of them were senior citizens. The youngest was one year old, and the oldest was 94 years old. The data's age distribution is skewed, indicating that the population with a low age is not there. A low-age population is not present, so the standard deviation is 19.1, indicating that the age group from 57 to 73 years old is not particularly densely populated. The normal level of TSH is between 0.5 and 5.0 mIU/L.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. EXPLORATORAY DATA ANALYSIS

An exploratory data analysis (EDA) utilizes data visualization to examine, investigate, and characterize the essential aspects of a data set. It is typically used to study what data may reveal outside of formal modeling or hypothesis testing tasks to better understand the relationships between components of the data set. Furthermore, it can help us to determine whether the statistical procedures we are considering for data analysis are appropriate. There are 27 attributes in our dataset.

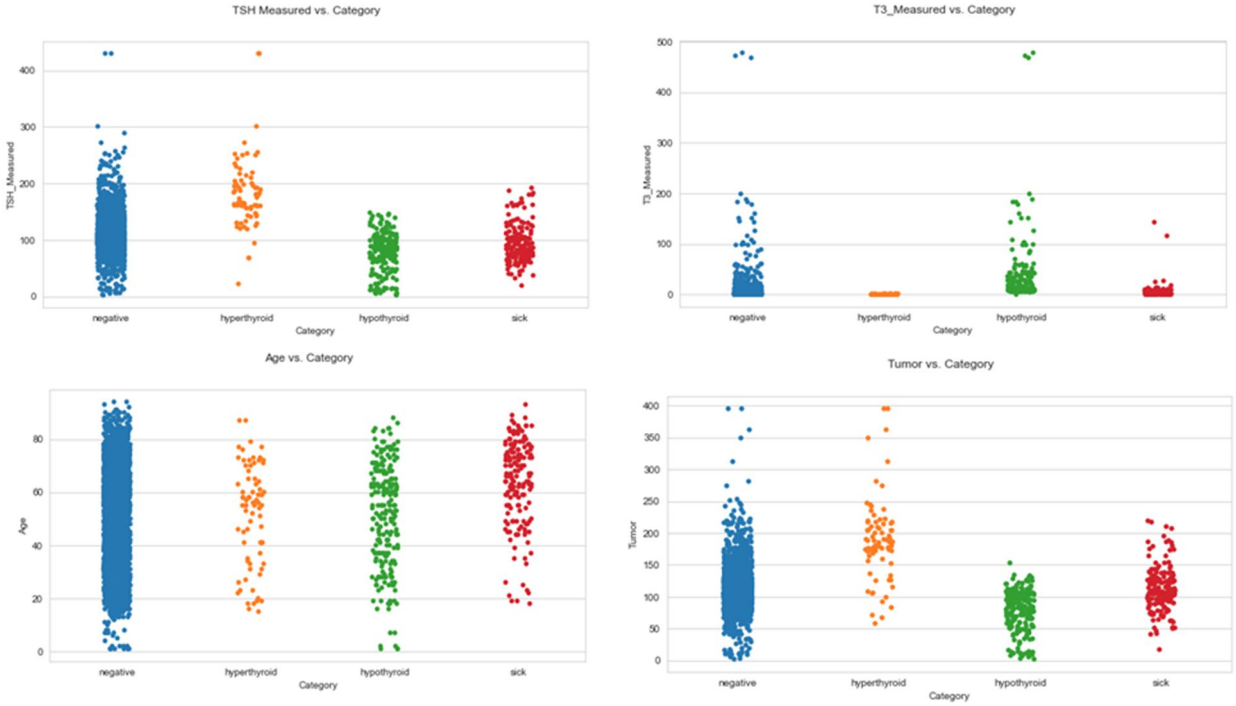


Fig. 2. Distribution chart of selected attributes

The figures above illustrate the distribution chart of various attributes based on the four target classes. When examining the cause-and-effect relationship between two variables, correlation is frequently used. A positive correlation occurs when two variables move in the same direction; as one increases, the other increases as well.

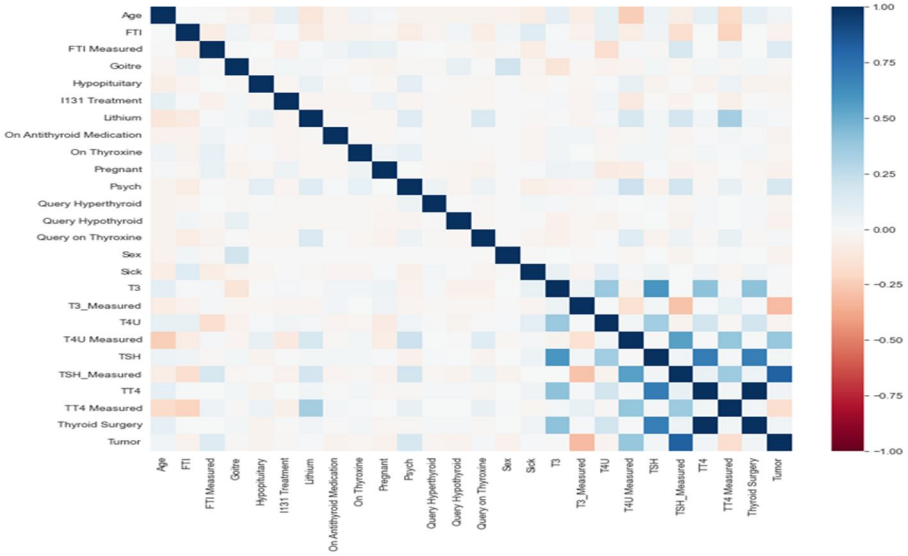


Fig. 3. Correlation Between Different Attributes From The Dataset

In the above figure, we can see the correlation between different attributes from the dataset. The above correlation shows that TSL_Measured and Tumor are highly correlated, whereas TT4 and Thyroid Surgery are correlated, respectively.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

V. BUILDING A MULTI-CLASS CLASSIFICATION MODEL

Two options are available for handling multiple classes in SKLearn: 'over' and 'multinomial'. They should be specified when initializing the Logistic Regression object under the 'multi_class=' option.

```
LogisticRegression
LogisticRegression(multi_class='ovr')
```

```
LogisticRegression
LogisticRegression(multi_class='multinomial')
```

OVR, which stands for One-Versus-Rest Heuristic, fits a binary regression for each label of your dependent variable by comparing the log-odds of that label with the combined log-odds of all the other labels. By selecting the multinomial option, a series of binary regressions is created in which each class label is compared with each of the other class labels individually. In the case of a dependent variable with k labels, OVR fits k number of models, while multinomial fits (k)*(k-1)/(2) number of models. In order to be explicit, consider a four-level dependent variable (Category: hyperthyroidism, hypothyroidism, sick and normal. An OVR would fit four models, and a multinomial model would fit six models.

Based on a set of covariates, predictions are made by estimating the probability of the outcome in each model. Other packages may offer voting options in order to assign each observation to a class based on a majority vote among the functions. The user is encouraged to investigate the scoring option. Regarding building inferential models, the biggest difference between these options is in the interpretation of coefficients. Exponentiated parameter estimates can be interpreted as the odds ratio associated with belonging to a modeled class compared to all other classes associated with a one-unit change in that parameter.

Multinomial logistic regression models typically define one of the k labels as a global referent class and then fit k-1 regressions, comparing the log-odds of each label to the global referent. Exponentiated parameter estimates are then interpreted as odds ratios associated with being in a modelled class compared to a global referent class with one unit change in that parameter.

VI. RESULT AND DISCUSSION

The multiclass classification of thyroid disease was analyzed using a logistic regression machine learning technique. Initially, the dataset contained 27 attributes and 3221 records. Comparisons of logistic regression models are made based on precision recall f1-score support. Logistic regression was used to estimate the regressor using two models: OVR and multinomial. In terms of accuracy, Logistic Regression outperforms. In spite of this, this algorithm does not have high precision, recall, or F1 scores. Our dataset is best predicted using Logistic Regression. Multiclass classification is attempted using Logistic Regression in this study. In order to estimate the concert of classifiers, the two models are trained with 75% of the whole data, and the remaining 25% are used to estimate the ensemble of classifiers. There are four classes of data in total. A logistic regression model will be the most appropriate model for classifying thyroid data into multiple classes. The figure below illustrates the results for two models.

	precision	recall	f1-score	support
hyperthyroid	0.00	0.00	0.00	19
hypothyroid	0.50	0.02	0.04	55
negative	0.86	1.00	0.92	689
sick	0.00	0.00	0.00	43
accuracy			0.85	806
macro avg	0.34	0.25	0.24	806
weighted avg	0.77	0.85	0.79	806

Fig. 4. Performance of Logistic regression with One-vs-Rest Heuristic

According to the above figures, the accuracy of Logistic regression with One-vs-Rest Heuristic is around 85%, which is a near-optimal model.

©IJRASET: All Rights are Reserved

	precision	recall	f1-score	support
hyperthyroid	1.00	0.05	0.10	19
hypothyroid	0.57	0.07	0.13	55
negative	0.86	0.99	0.92	689
sick	0.67	0.05	0.09	43
accuracy			0.86	806
macro avg	0.77	0.29	0.31	806
weighted avg	0.83	0.86	0.80	806

Fig. 5. Performance of Logistic regression with Multinomial

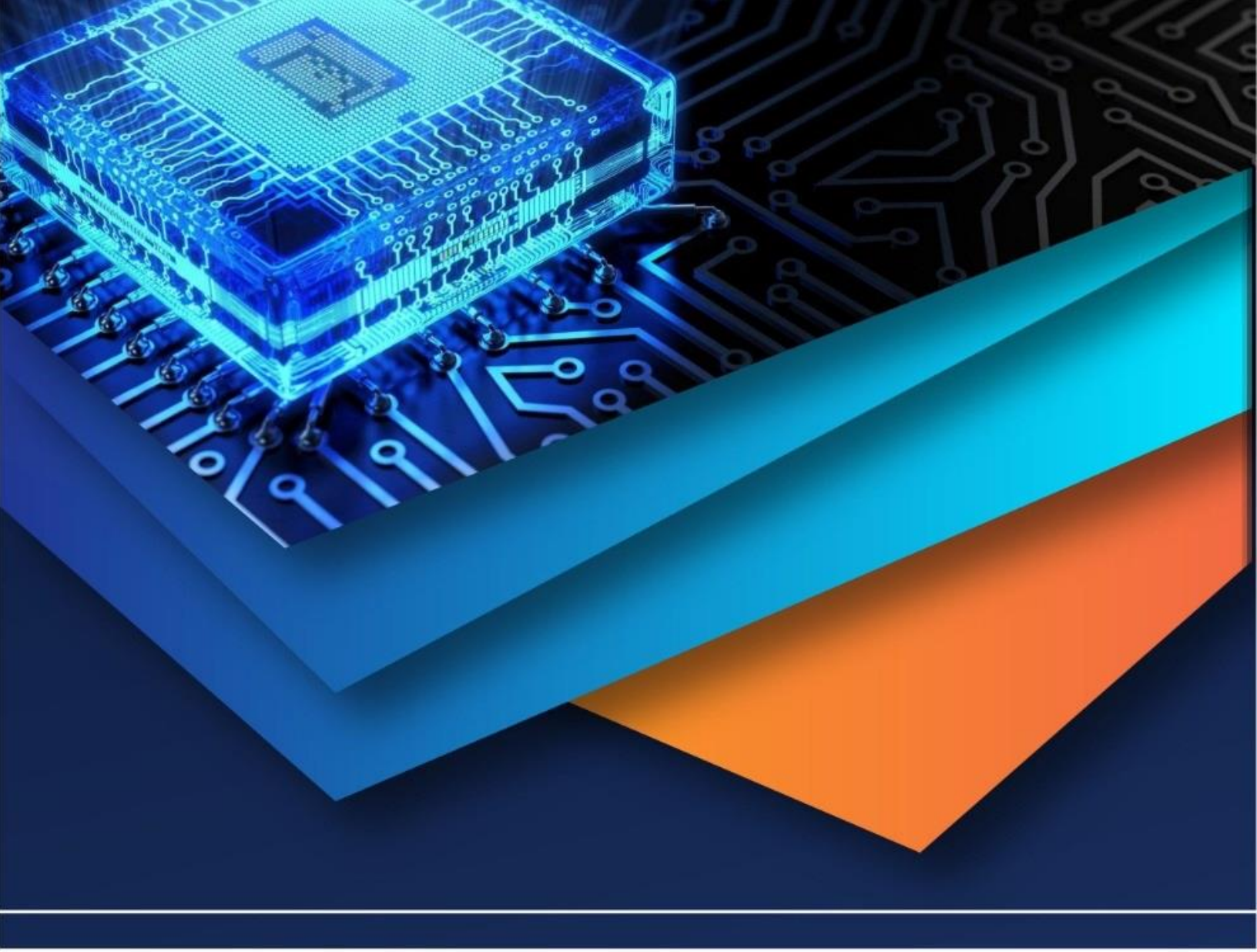
It can be seen from the above figures that the accuracy of Logistic regression with Multinomial is approximately 86%, which is significantly better than the One-vs-Rest Heuristic.

VII. CONCLUSION

In order to classify thyroid disorders using ML techniques, logistic regression was examined for four-class classification and two predictors have been used, OVR and Multinomial, to estimate the regressor. Precision, recall, F1-score, and accuracy have been used to compare the performance of predictive models. Prediction models built using logistic regression with multinomial provide the highest accuracy of 86%, while those built with the One-vs.-Rest Heuristic result in an accuracy of 85%. As a result of these algorithms, in the future, more real-time data related to thyroid disease can be used for the prediction of thyroid disease. This work may be used in hospitals to assist doctors and clinicians in diagnosing hypothyroidism. It is also possible to increase the dataset, and academic researchers in the medical field can use this work to identify more ML-based prediction models for classifying thyroid disease.

REFERENCES

[1] Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V., “Thyroid Disease Prediction Using Machine Learning Approaches”, The National Academy of Sciences, 2020.
[2] Rao, A. R., & Renuka, B. S.,“A Machine Learning Approach to Predict Thyroid Disease at Early Stages of Diagnosis”, IEEE International Conference for Innovation in Technology (INOCON). Bangluru, India: IEEE, 2020.
[3] Aversano, L., Bernardi, M. L., Cimitile, M., Iammarino, M., Macchia, P. E., Nettore, I. C., & Verdone, C.,“Thyroid Disease Treatment prediction with machine learning approaches”, Procedia Computer Science, 1031-1040, 2021.
[4] Chandan, R., Vasan, C., Chethan , M., & Devikarani , H.,“THYROID DETECTION USING MACHINE LEARNING”, International Journal of Engineering Applied Sciences and Technology, 173-177, 2021 .
[5] More, K., “Classification of Thyroid Disease using Machine Learning” International Research Journal of Engineering and Technology (IRJET), 261-265, 2021.
[6] Raihan Asif, M.-A., Mirza , M. N., Faisal, F., Shikder, M., Udoy, M. H., & Ahsan, R., “Computer Aided Diagnosis of Thyroid Disease Using Machine Learning Algorithms”, 2020 11th International Conference on Electrical and Computer Engineering (ICECE). Dhaka, Bangladesh: IEEE, 2021.
[7] Rehman, H. A., Lin, C.-Y., Mushtaq, Z., & Su , S.-F., “Performance Analysis of Machine Learning Algorithms for Thyroid Disease”, Arabian Journal for Science and Engineering, 9437–9449, 2021.
[8] Tahir , A., Muhammad , H., Khalid , A., Tauqeer, F., Nadia , T., & Aqeel , A., “Empirical Method for Thyroid Disease Classification”, BioMed Research Internationa, 10, 2022.
[9] Verma, S., Popli, R., Kumar, H., & Atul, S. “Classification of thyroid diseases using machine learning frameworks”, International Journal of Health Sciences, 7552–7566, 2022.
[10] Zahrul , J. P., Chumki, M. N., & Khan , M. Z., “Predictive Analysis for Thyroid Diseases Diagnosis Using Machine Learning”, 2021 International Conference on Science & Contemporary Technologies (ICSCT) Dhaka, Bangladesh: IEEE, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)