Performance Analysis of Machine Learning Algorithms in Thyroid Disease Prediction

Dr. E. Haripriya
Department of Computer Science, J. K. K. Nataraja College of Arts & Science, Namakkal, India

Abstract: In Health Care Systems to deal with the huge amount of data, Machine Learning algorithms play the vital role for earlier disease detection and prediction. In recent years, the most common issue identified in many women and men in their adolescent age was Thyroid issues. The issue once detected cannot be cured but the symptoms can be managed by proper treatment. Thyroid can be identified by the symptoms such as on thyroxine, query on thyroxine, query hypothyroid etc. Early identification and diagnosis are the basic factor for the correct treatment. In our study, the performance of the four Machine Learning Algorithms Decision Tree Classifier, Support Vector Machine, Random Forest Regressor and KNeighbors Classifier are analyzed and compared to predict the disease based on the performance metrics such as accuracy, recall and precision.

Keywords: *Machine Learning, Thyroid, Decision Tree, Support Vector Machine, Random Forest Regressor, KNeighbors Classifier.*

I. INTRODUCTION

In Health care Industry, Computational Biology plays a vital role for predicting the disease using the stored patient's data. Here, for predicting the disease in early stage, machine learning algorithms are used. Machine learning models uses the features selected from several datasets for prediction as accurate as possible. Since proper classification is necessary to treat the affected patients and to avoid unnecessary treatments for healthy patients.

The thyroid gland is found at the front of the neck under the voice box and plays a major role in the metabolism, growth and development of the body. Most of the body functions are regulated by thyroid hormones by constant release of standard amount of thyroid hormones into the blood stream. The hormones produced by thyroid gland are 1. Triodothyronine (T3) 2. Tetraiodothronine (T4). Iodine is the major building blocks of above hormones. Since the human body didn't produce this element through diet only enough amount of Iodine is taken. If the thyroid gland produces too many hormones it leads to hyperthyroidism and if the gland doesn't make enough hormones it leads to hypothyroidism.

II. LITERATURE REVIEW

Andita et.al compared the performance of the classification algorithms such as support vector machine, Artificial Neural Network, K-Nearest Neighbor. Here the best algorithm is found by analyzing the accuracy measure while predicting the target output.

Sunila et.al compared the performance of the Logistics Regression and Support Vector Machine algorithms. Here the performance of the machine learning algorithms are compared based on the precision, recall, I measure and RMS error.

Khalod et.al compared the performance based on the accracy measure of Support vector machine, Random Forest, Decision tree, Naïve Bayes and K-Nearest Neighbor. The dataset is collected from external hospitals and laboratories over a period of four months and it contains 17 features.

Saima et.al examined multiple machine learning algorithms including CatBoost, Artificial Neural Network, Light GBM, KNN to improve the prediction accuracy. The performance of the machine learning model is analyzed based on the accuracy, prediction, recall and F1 scores.

Priyanka et.al compared the performance of Naïve Bayes, Support Vector Machine and Random Forest. Here the resultant set is classified as 4 classes named Hypo, Hyper, Sick Euthyroid & Euthyroid. They used the Univariate feature selection, Recursive and Feature Elimination & Tree based feature selection for selecting the dataset attributes.

III. PROBLEM DEFINITION

Thyroid disorders increasing day by day in India and for women in their pregnancy period suffer a lot due to lack of prediction in the earlier stages. Prediction of thyroid disorder in kids and women is a complicated process for a doctor. Two or more Expert's opinion is necessary for the proper prediction. The main objective of the proposed system is to develop a model that can be used to predict the thyroid disease more accurately with minimal features. To help doctors in earlier prediction and diagnosis, the machine learning algorithms provides necessary assistance and reduce their burden.

IV. PROPOSED SYSTEM

Analyzing the blood report dataset is necessary to predict the disease and various ML algorithms are used for categorizing the thyroid disorders and the best algorithm will be chosen based on the accuracy, precision and recall values. The dataset contains the features of the patients having hypo and hyperthyroidism. Data preprocessing and cleaning is done to fill the null values or Nan values before splitting into training and testing set. Preprocessed data is used as the training and testing dataset. Feature scaling is done using standard scaler for equal contribution of the attributes in target value prediction. After scaling, the dataset is given as the input to the algorithm. Model is developed using python where the chosen ML model predict the target value. The users blood test data can be entered for custom prediction and the application will process the data using the model and result will be displayed on the screen.

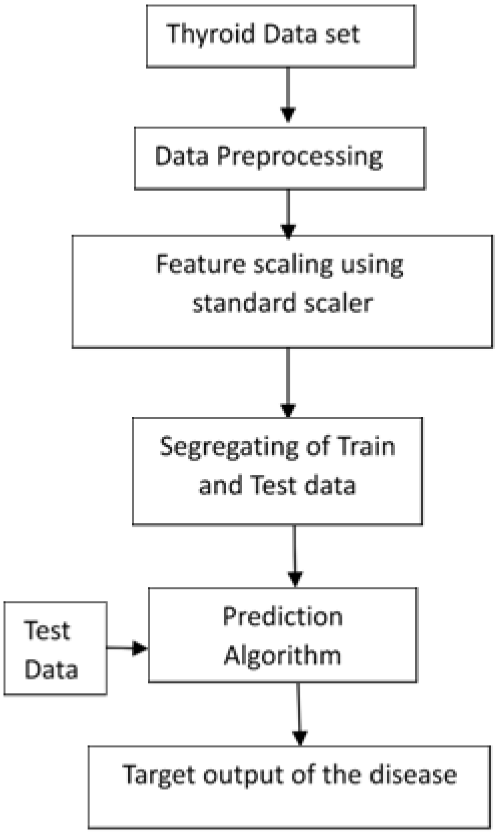

Thyroid Data set
Data Preprocessing
Feature scaling using standard scaler
Segregating of Train and Test data
Test Data
Prediction Algorithm
Target output of the disease

Fig 1: Flow chart of the proposed system

V. RESULTS AND DISCUSSION

Data Set

A new thyroid dataset is taken from Kaggle repository and it has 1000 instances and 27 features. Lab test are taken to predict thyroid. Out of these 27 features 7 features such as on thyroxine, query hypothyroid, TSH measured, TSH, T3, TT4 and FTI are selected using correlation matrix with correlation target greater than 0.07.

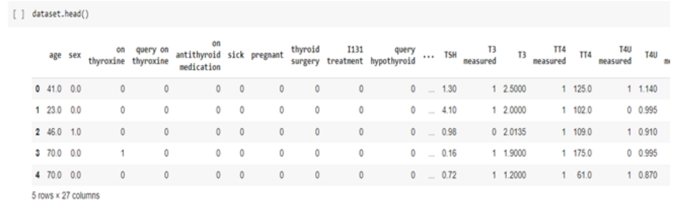
Table 1: Dataset

Dataset	No. of Features	No. of Instances	No.of Classes
Thyroid	27	1000	2

Table 2 : Thyroid Dataset Attributes

S.No	Attribute Name	Data Type
1	age	float64
2	Sex	float 64
3	on thyroxine	int64
4	Query on thyroxine	int64
5	Antithyroid medication	int64
6	Sick	int64
7	Pregnant	int64
8	thyroid surgery	int64
9	I131 treatment	int64
10	query hypothyroid	int64
11	Lithium	int64
12	Goitre	int64
13	Tumor	int64
14	Hypopituitary	int64
15	Psych	int64
16	TSH measured	int64
17	TSH	float64
18	T3 measured	int64
19	T3	float64
20	TT4 measured	int64
21	TT4	float 64
22	T4U measured	int64
23	T4U	float64
24	FTI measured	int64
25	FTI	float64
26	binaryClass	int64

Table 3: Dataset Header



```
[ ] dataset.head()
```

```
                            on                              on                                                                            T3             TT4             T4U
   age  sex    on  query on antithyroid sick pregnant thyroid   I131       query  ...   TSH  measured     T3  measured   TT4  measured    T4U  m
             thyroxine thyroxine medication               surgery treatment hypothyroid
0  41.0  0.0     0        0        0      0       0       0        0          0  ...  1.30        1  2.5000        1  125.0       1  1.140
1  23.0  0.0     0        0        0      0       0       0        0          0  ...  4.10        1  2.0000        1  102.0       0  0.995
2  46.0  1.0     0        0        0      0       0       0        0          0  ...  0.98        0  2.0135        1  109.0       1  0.910
3  70.0  0.0     1        0        0      0       0       0        0          0  ...  0.16        1  1.9000        1  175.0       0  0.995
4  70.0  0.0     0        0        0      0       0       0        0          0  ...  0.72        1  1.2000        1   61.0       1  0.870
5 rows × 27 columns
```

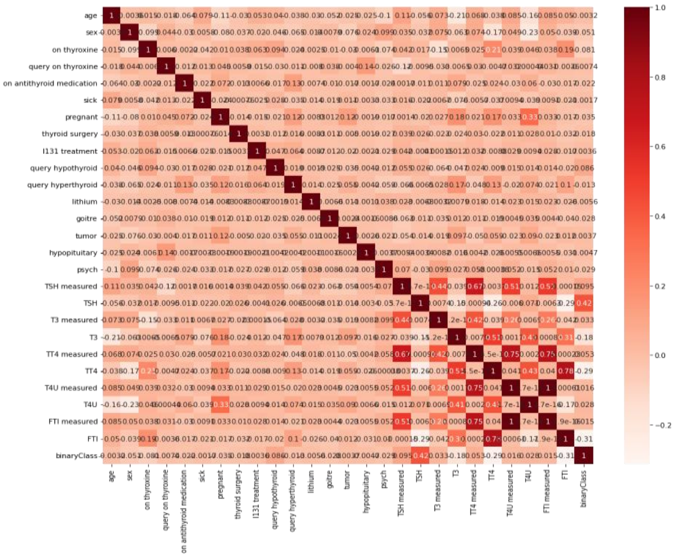
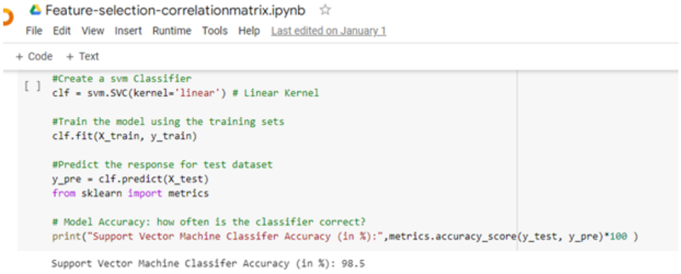
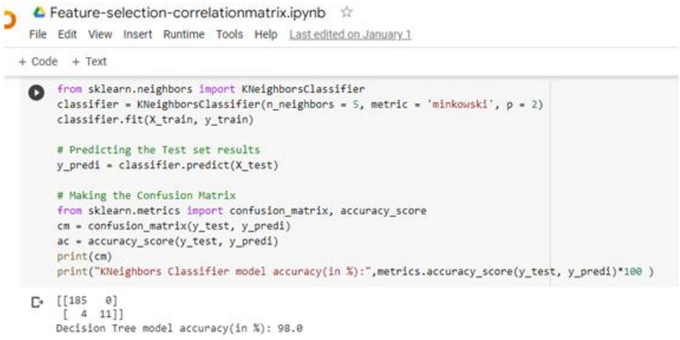


Fig 2 : Screenshot of Feature Selection using Correlation Matrix



```
[ ] #Create a svm Classifier
    clf = svm.SVC(kernel='linear') # Linear Kernel

    #Train the model using the training sets
    clf.fit(X_train, y_train)

    #Predict the response for test dataset
    y_pre = clf.predict(X_test)
    from sklearn import metrics

    # Model Accuracy: how often is the classifier correct?
    print("Support Vector Machine Classifer Accuracy (in %):",metrics.accuracy_score(y_test, y_pre)*100 )

    Support Vector Machine Classifer Accuracy (in %): 98.5
```

Fig 3: Support Vector Machine Accuracy


Feature-selection-correlationmatrix.ipynb
File Edit View Insert Runtime Tools Help Last edited on January 1
+ Code + Text

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
classifier.fit(X_train, y_train)

# Predicting the Test set results
y_predi = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_predi)
ac = accuracy_score(y_test, y_predi)
print(cm)
print("KNeighbors Classifier model accuracy(in %):",metrics.accuracy_score(y_test, y_predi)*100 )
```

```
[[185   0]
 [  4  11]]
Decision Tree model accuracy(in %): 98.0
```

Fig 4 : K-Neighbors Classifier Model Accuracy

Comparison of Models

The following observations were obtained after implementing the machine learning models in Python notebook using sklearn library. The Decision Tree classifier shows the accuracy prediction as 99% and the recall value as 0.933. The Random Forest Regressor has the accuracy value as 92%. The KNeighbors Classifier shows the accuracy prediction as 98% and the Precision and recall values as 1 and 0.733 respectively. The Support vector machine Classifier shows the accuracy prediction as 98.5% and the precision and recall values as 1 and 0.8 respectively. The Decision tree classifier outperforms other ML Algorithms while considering the accuracy score and recall value to predict the thyroid disease.

CONCLUSION

The machine learning algorithm usage in predicting thyroid disease is a simple and smart approach. Here four machine learning models are used and compared and among those models decision tree classifier prediction is more accurate than other model. Custom prediction is also done by giving the blood test samples of the user as the input. The main objective of this model is to predict the disease as earlier as possible and give a proper guidance and smart approach to the society.

References

[1] Chandran R, Chetan Vasan, Chethan MS &Devikarani HS, "Thyroid detection using machine learning", International Journal of Engineering Applied Sciences and Technology, 2021, Vol.5, Issue.9, pp.173-177.

[2] Ankita Tyagi and Ritika Mehra, "Interactive Thyroid Disease prediction system using machine learning technique", Fifth International Conference on Parallel, Distributed and Grid Computing, 2018.

[3] SunilaGodara& Sanjeev Kumar, "Prediction of Thyroid Disease Using Machine Learning Techniques", International Journal of Electronics Engineering, 2018, Vol.10, No.2,pp. 787-793.

[4] Saima Sharleen Islam, Samiul HaqueMd, Saef Ullah Miah M, Talha Biu Sarvar & Ramdhan Nugraha, "Application of machine learning algorithms to predict the thyroid thyroid disease risk: an experimental comparative study", Peer Computer Science, 2022, Vol.8, pp.1-35.

[5] Priyanka Duggal & Shipra Shukla, "Prediction of Thyroid Disorders using Advanced Machine Learning Techniques", IEEE 10th International Conference on cloud computing, Data Science and Engineering.

[6] Khalid salman, Emruallah Sonic, "Thyroid Disease Classification using Machine learning Algorithms", Journal of Physics, 2021, pp.1-13.