A FRAMEWORK FOR THYROID DISEASE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

[1]Dr. N. Baggyalakshmi, [2]Dr. R. Revathi, [3]Dr. Bosco Paul Alapatt, [4]Dr. Felix. M. Philip

[1]Assistant Professor, Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore - 21.
baggyanethra@gmail.com
[2]Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore - 21.
revathilakshay@gmail.com
[3]Assistant Professor, School of Sciences, CHRIST (Deemed To Be University), NCR Campus, Delhi.
bosco.paul@christuniversity.in
[4]Assistant Professor, Department of Computer Science, Jain (Deemed To Be University) Kochi.
m.felix@jainuniversity.ac.in

Abstract—The Thyroid gland is a vascular gland and one of the most important organs of a human body. This gland secretes two hormones which help in controlling the metabolism of the body. Thyroid disease is a major cause of formation in medical diagnosis and in the prediction, onset to which it is a difficult axiom in the medical research. The two types of Thyroid disorders are Hyperthyroidism and Hypothyroidism. When this disorder occurs in the body, they release certain type of hormones into the body which imbalances the body's metabolism. Thyroid related Blood test is used to detect this disease but it is often blurred and noise will be present. Data cleansing methods were used to make the data primitive enough for the analytics to show the risk of patients getting this disease. The machine learning plays a decisive role in the process of disease prediction and this paper handles the analysis and classification models that are being used in the thyroid disease based on the information gathered from the dataset taken from UCI machine learning repository. In this paper few machine learning techniques for diagnosis and prevention of thyroid.

Keywords: Thyroid Disease, Naïve Bayse, kNN, Decision Tree, Machine Learning Algorithms.

I. Introduction

At least a person out of ten is suffered from thyroid diseasein India. The disorder of thyroid disease primarily happensin the women having the age of 17 to 54. The evolvement computational biology is used in healthcare industry. It allows collection of stored patient data for the prediction of the disease. There are prediction algorithms which are available for the diagnosis of the disease at early stages. The medical information systems are rich of datasets but there are only few intelligent systems which can easily analysis the disease. In any disease prediction models are used to override the features that can be selected from different datasets which can be used in classification in healthy patient as accurate as possible [1]. If this is not done, misclassification can lead to a healthy patient getting unnecessary treatment. The Thyroid gland is an endocrine gland present in the human neck beneath the Adam's apple which help in secretion of thyroid hormone that influence the rate of metabolism and protein synthesis. The thyroid hormones are useful in counting how briskly the heart beats and how fast we burn calories. The thyroid secretes two types of active hormones called Levothyroxine (T4) and Triiodothyronine (T3) [2]. There are various kinds of medications like thyroid surgery is liable to ionizing radiation, continual tenderness of the thyroid, deficiency of iodine and lack of enzyme to make thyroid hormones.

Cure of disease is a regular concern for the health care practitioners, and the errorless diagnostic at the right time for a patient is very important. Recently, by some advanced diagnosis methods, the common medical report can be generated with an additional report based on symptoms.

Health care data can be processed and after implementing with certain methodologies; it can provide information that can be used in diagnosis and treatment of diseases more efficiently and accurately with better decision making and minimizing the death risk. For diagnosis entire medical history and physical tests (T4, T3Test, Cholesterol test, TSH Test) are required[8]. As these test produces large amount of data and ML can be used for finding important features from large amount of data. Due to this specialty of ML can be used in combination with medical science for the accurate diagnosis of hypo thyroid disease[9].

2. Problem Definition

According to statistics, thyroid disorders are on the rise in India. Approximately 1 in 10 Indian adults suffer from thyroid problem. It has been estimated that around 42 million peoples suffer from thyroid disease. At least a person out of ten is suffered from thyroid diseasein India. To assist doctors machine learning can help them in diagnosis of disease and reduces their burden [3]. The main objective is to develop a system which can predict the type of thyroid disease that patient is affected from. To predict thyroid disease with usage of minimum number of parameters.

3. Research Methods

For predicting Thyroid disease analyzing blood report is required to analyze and predict disease. Thyroid blood test data set analysis will be conducted using various supervised machine learning classifier techniques. Based on the accuracy of different algorithm, best accuracy algorithm will be chosen to fetch the result.

3.1 Data Collection

For first part, thyroid data set is taken from UCI repository[6]. The dataset of hypothyroidism is used where negative and hypos are the two labels.

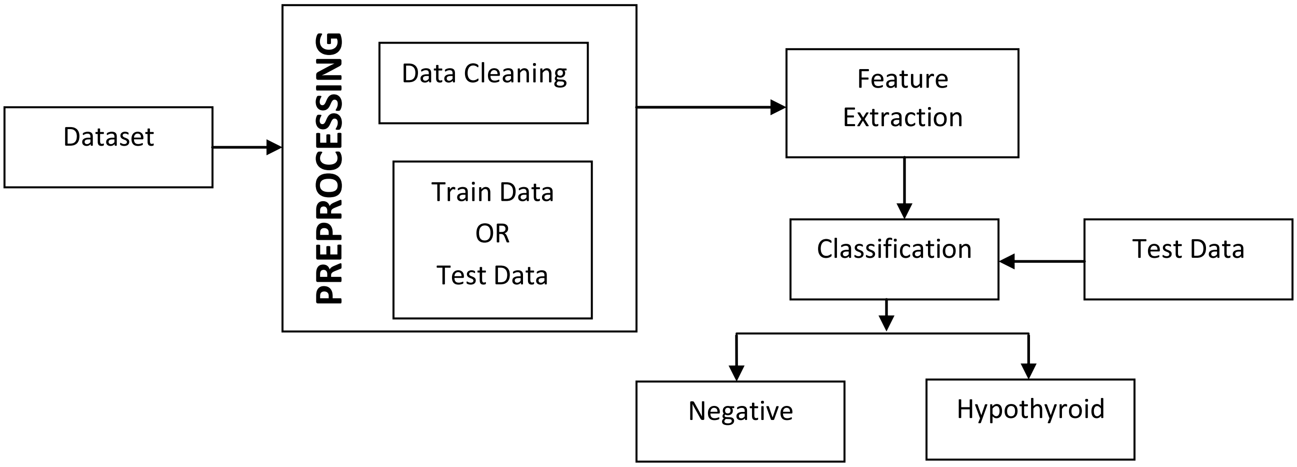
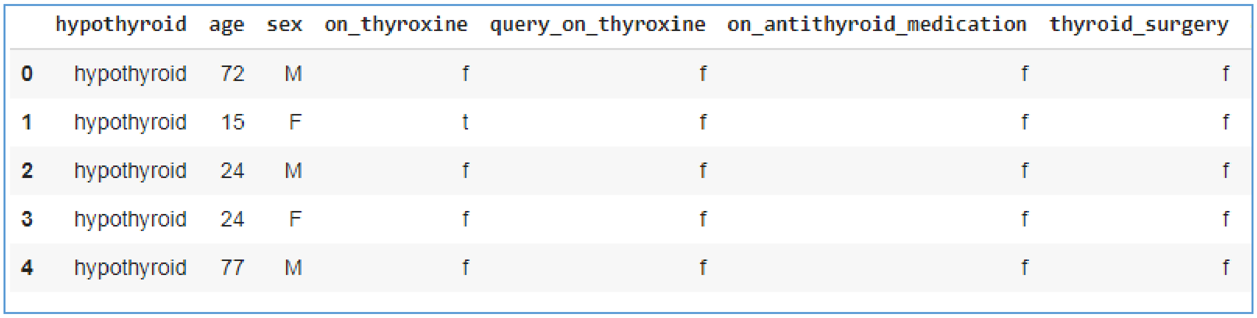


Fig: 1 Proposed Architecture for Thyroid classification system

3.2 Dataset: A thyroid dataset accessed from UCI repository. This set has 3163 samples and 26 columns and sample entries were shown in figure 2. Dataset are used to try to predict whether a patient's thyroid to the class is normal, hypothyroidism or hyperthyroidism[10]. The diagnosis at the class level was based on a complete medical record.

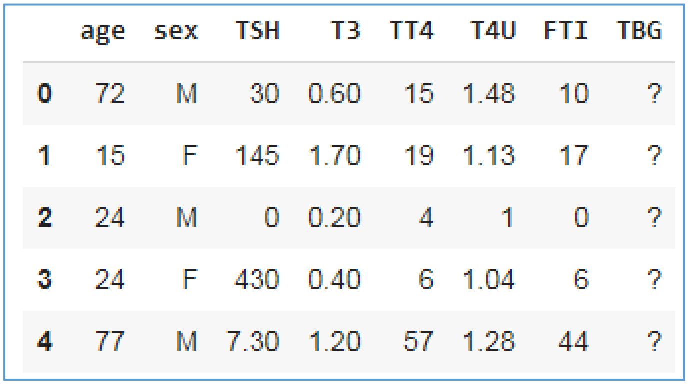


	hypothyroid	age	sex	on_thyroxine	query_on_thyroxine	on_antithyroid_medication	thyroid_surgery
0	hypothyroid	72	M	f	f	f	f
1	hypothyroid	15	F	t	f	f	f
2	hypothyroid	24	M	f	f	f	f
3	hypothyroid	24	F	f	f	f	f
4	hypothyroid	77	M	f	f	f	f

Fig: 2 Sample Dataset

3.3 Data Cleaning

The data set need to be checked before feeding it to training. There may be presence of null data or unnecessary data; this should undergo data cleaning to remove such data.Cleaned data is used as training data and test data, which is fed as input to the algorithm.Missing values, non numeric valuesreplaced with proper data values as shown in figure 3. kNNImputer data preprocessing algorithm is used for replacing the missing values. It is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset[7]. It replaces the NaN values with a specified placeholder. Shifted Class attributes to the last column for further usage.

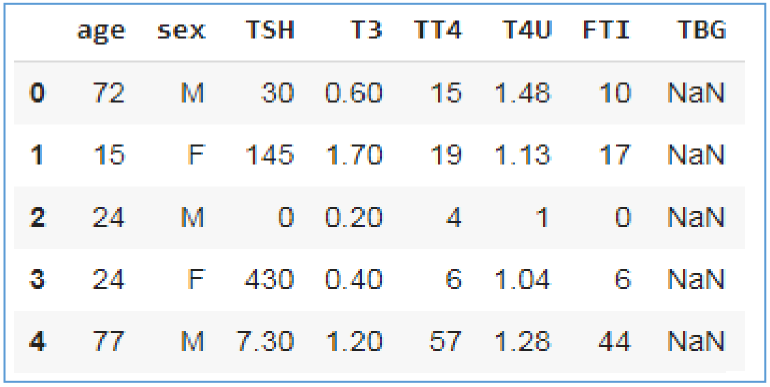


	age	sex	TSH	T3	TT4	T4U	FTI	TBG
0	72	M	30	0.60	15	1.48	10	?
1	15	F	145	1.70	19	1.13	17	?
2	24	M	0	0.20	4	1	0	?
3	24	F	430	0.40	6	1.04	6	?
4	77	M	7.30	1.20	57	1.28	44	?

Fig: 3 Noisy Data Values present in the dataset

3.4 Feature Extraction

The algorithm extracts the features from different dataset to classify the data according to the labels. To check the accuracy of the prediction, test data is fed to the algorithm[4].Based on the feature extracted, probability will be generated for test data by comparing the features of both. Highest probability value will be classified to that particular label whether it is hypothyroid or negative[13]. Age, Sex, TSH, T3, TT4, T4U, FTI, and TBG used to figure out the features as shown in figure 4.



	age	sex	TSH	T3	TT4	T4U	FTI	TBG
0	72	M	30	0.60	15	1.48	10	NaN
1	15	F	145	1.70	19	1.13	17	NaN
2	24	M	0	0.20	4	1	0	NaN
3	24	F	430	0.40	6	1.04	6	NaN
4	77	M	7.30	1.20	57	1.28	44	NaN

Fig: 4 Visualize Features

To handle categorical feature 'sex' column, will use frequent category imputation, After using that to ‘sex’ attribute it is plotted in bar chart as shown in figure 5. Next process is to handle categorical features. To do so have to create a separate variable with categorical features present inside our dataset.

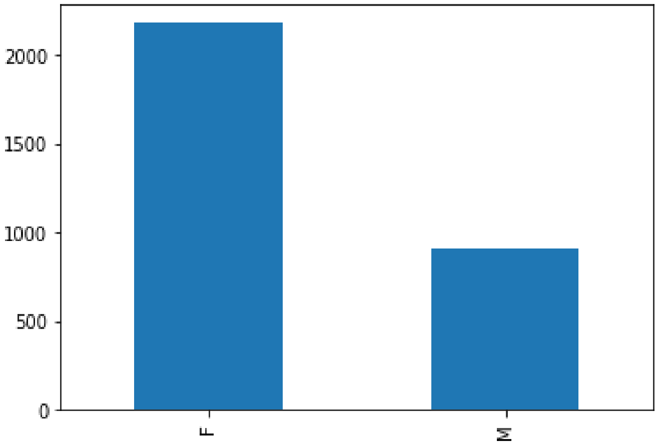


Fig: 5 ‘Sex’ Attribute

After assigning separate variable, we can observe that all categorical features have only two categories. There after label encoding technique used to reduce the complexity. Label encoded as such M--1, F--0, f--0, t--1, y--1, n--0, hypothyroid--0, negative—1[6].

Distribution of all the variables before doing feature selection is highly essential (as shown in fig 6). Histogram representation used for distribution. Few observations from the histograms:

1. age, T3, TT4,T4U,FTI, are following normal distribution (nearly).
2. Rest all variables are categorical variables, so there is no concept of normal distribution

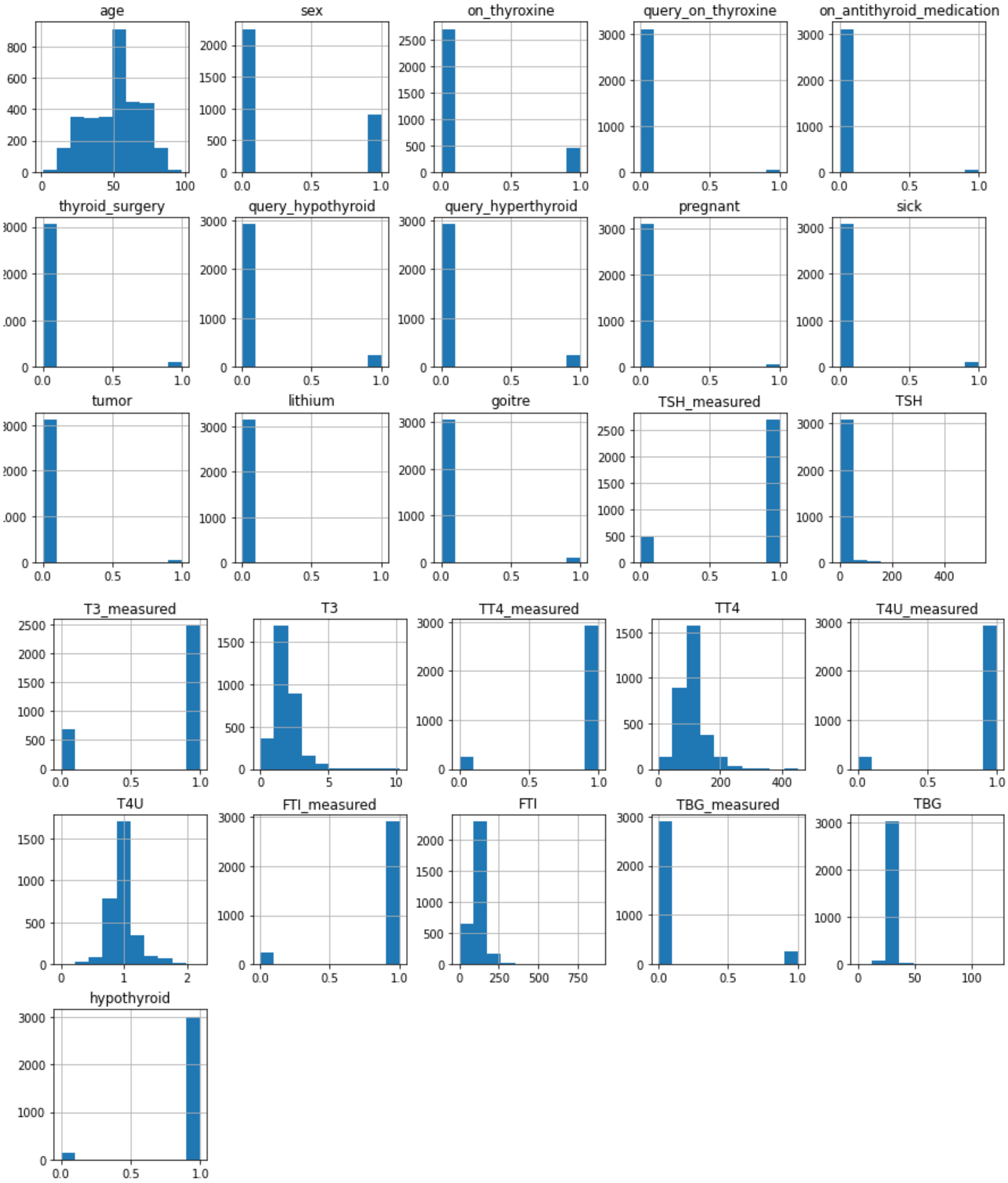


Fig: 6 Histogram for Variable Distribution

3.5 Classification:

3.5.1 Naïve Bayes

Naïve Bayes is a probabilistic classification method that uses Bayes theorem. The Naïve Bayes classifier takes a set of features from a dataset and determines the probability of each feature occurring in each class within the data [12] as shown in figure 7.

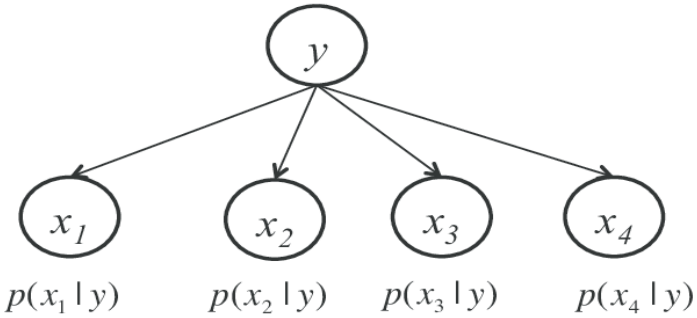


Fig: 7 Naïve Bayes

For each row of data, the values of the attributes are used to calculate the posterior probability for each class within the dataset, the row of data is then assigned to the class with the highest posterior probability. This method is referred to as naïve because it assumes that all features of the dataset are independent of one another, which is an assumption that is likely untrue and thus naïve.

The advancement in the Bayesian theory gets the evolution of Naïve Bayes algorithm. The Naïve Bayes is a supervised machine learning algorithm based on the Bayes Theorem [11]. The Bayes theorem for the likelihood is given as (1):

$$P\left({}^{Y}/_{X}\right) = \frac{p\left({}^{X}/_{Y}\right) * p(Y)}{p(X)} \tag{1}$$

Since in (1) the p(x) is constant and add extra calculation in the computation, hence it is being removed from the formula, and given as (2):

$$P\left({}^{Y}/_{X_i}\right) = \sum_i p\left({}^{X_i}/_{Y}\right) * p(Y) \tag{2}$$

3.5.2 k-NN:

The k-nearest neighbor (KNN) algorithm assigns class labels to rows within a dataset based on the class labels of training data that are similar [10]. The KNN algorithm works by searching the training data for k training tuples that are closest to the test data tuple and assigns the test tuple a class label based on the class labels of those closest training tuples (as shown in fig 8). The closeness of a training tuple to a test tuple is determined using a distance function, such as Euclidean distance.

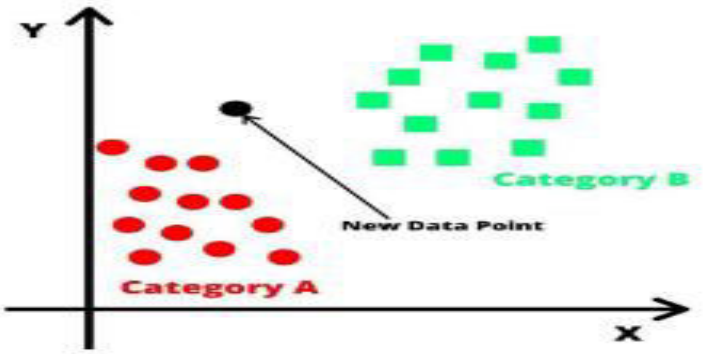


Fig: 8 Classification of dataset using k-NN

A refinement of the k-NN classification algorithm is to weigh the contribution of every ofthe k neighbors consistent with their distance to the query point xq, giving greater weight w_i to closer neighbors [5]. It is given by

$$f\left(x_q\right) = \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i} \tag{3}$$

Where the weight is,

$$w_i = \frac{1}{d(x_q, x_i)^2} \tag{4}$$

In case x_q exactly matches one among xi in order that the denominator becomes zero, weassign f(x_q) equals f(x_i) during this case. It is sensible to use all training examples not just k ifweighting is employed, the algorithm then becomes a worldwide one[14]. The main disadvantage isthat the run time of this algorithm is bit longer.

3.5.3 Decision tree

A decision tree is a structure that contains internal nodes that denote attributes, branches that denote the outcome of a test on an observation and leaf nodes that denote the class label [11]. The top node of this tree-like structure is the root node. In order to determine the class of an observation,(as shown in fig 9) the decision tree is followed, starting at the root, moving down to the leaf nodes.

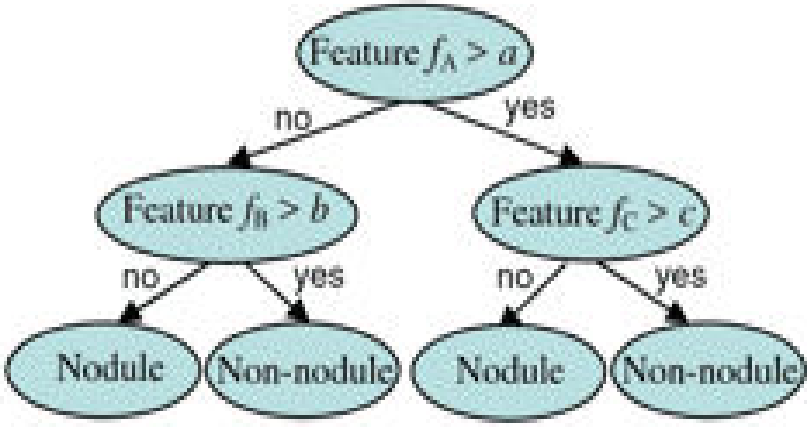


Fig: 9 Logical representation of Decision Tree

The decision trees are ways to find the conclusion based on the set of rules drawn from the tree. A decision tree is consists of the two nodes: i) Decision node and ii) Leaf node. A decision node tells about which attribute have to be selected and leaf nodetells about the class.Decision trees use the up down approach to give the results [11]. The first node of the decision trees, a decision node, called as root node [12]. The two important formulas that are used in this method are: i) Entropy calculation and ii) Information gain, for calculating the entropy of the sample data[10].

$$E(s) = \sum -p_i \log_2 p_i \tag{5}$$

After calculating the entropy, the information gain is calculated for each attribute to get decide thedecision node.

$$Gain(s,a) = Entropy(s) - \sum_{vEvalues(a)} |s|/|s| Entropy(s_{v)} \tag{6}$$

3.5.4 Confusion Matrix

To evaluate the performance of classification models confusion matrix [15] is used. It can be used for binary as well as multinomial classification. The target variable has positive and negative values represented in matrix as true positive (TP), true negative (TN), false positive (FP), false negative (FN). TP refers that the model predicted positive and actual was positive, in TN the model predicted negative and actual was negative, in FP the model predicted positive but actual was negative. FP is also known as type 1 error. In FN the model predicted negative but actual was positive. FN is also known as type 2 error.

4. Result and Discussion

4.1 Heat Map

A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. Correlation between parameters of our dataset is interpreted and pictorial view is obtained as shown in fig 10.

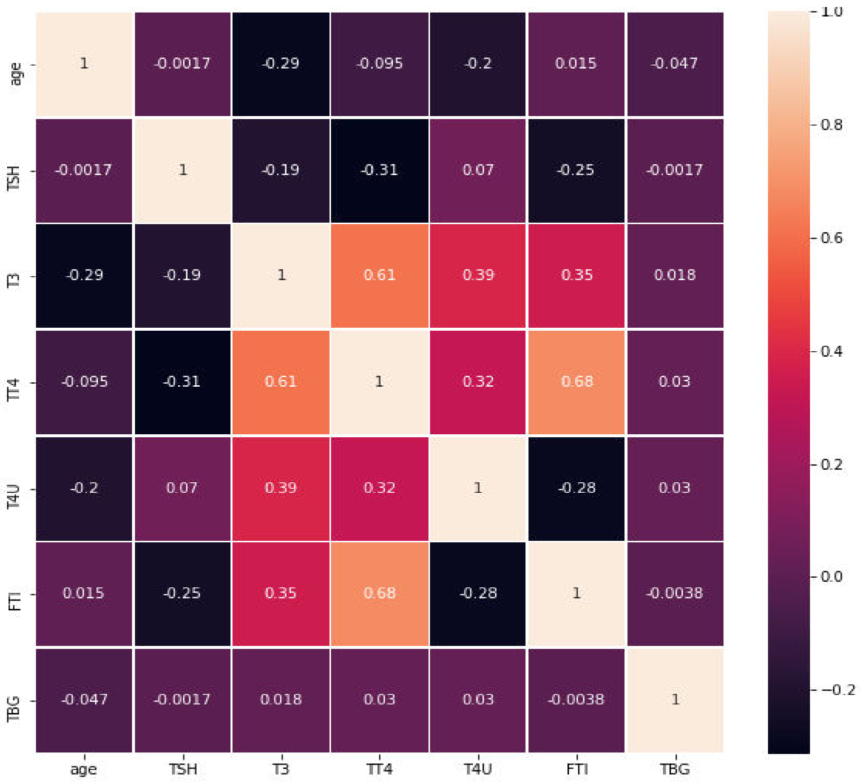
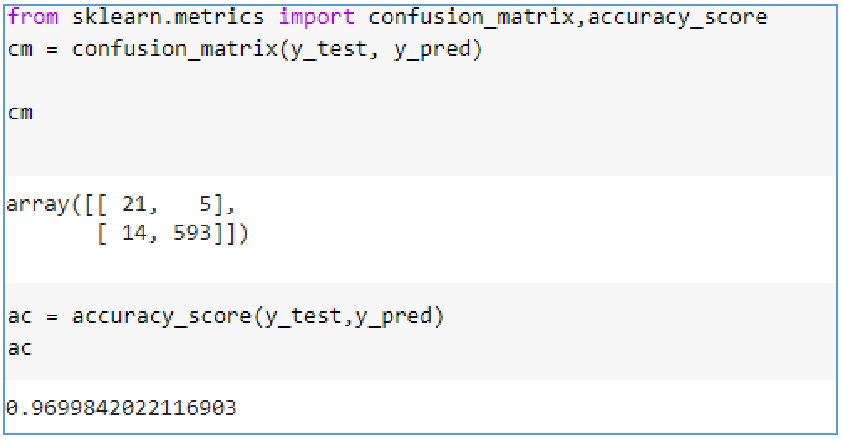


Fig: 10 Correlation representation using Heat Map

Naive Bayes (GaussianNB) algorithm has been implemented and 96.99% accuracy score is obtained as shown in figure 11.



```
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(y_test, y_pred)

cm
```

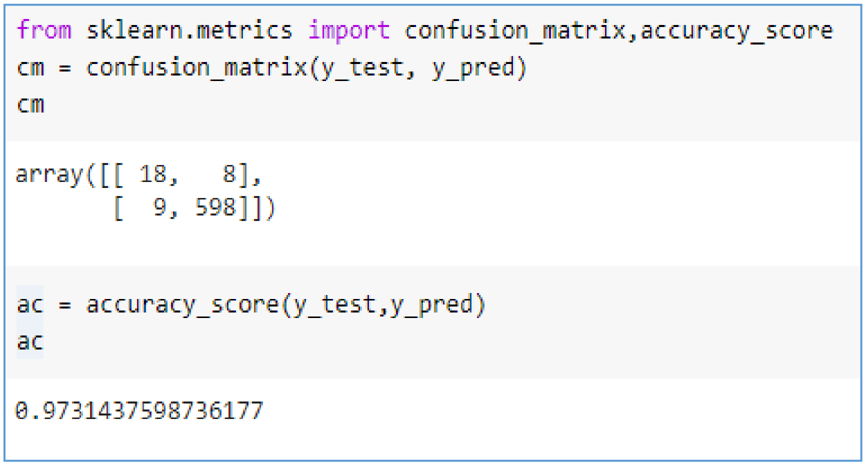
```
array([[ 21,   5],
       [ 14, 593]])
```

```
ac = accuracy_score(y_test,y_pred)
ac
```

```
0.9699842022116903
```

Fig: 11 Accuracy obtained using Naïve Bayes

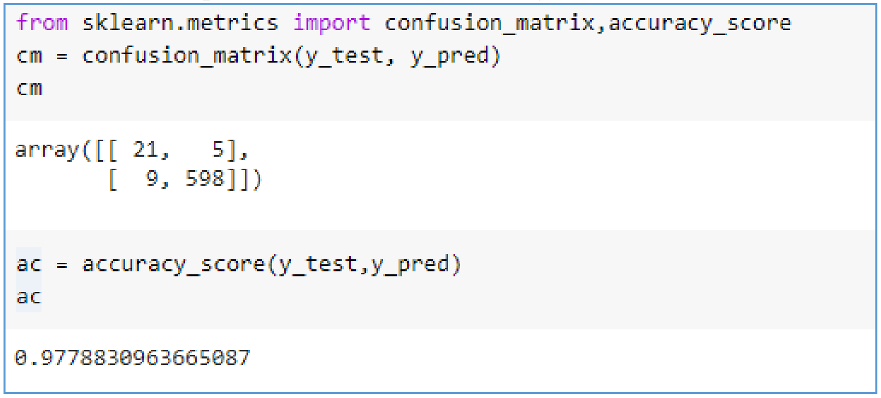
KNN algorithm has been implemented and 97.31% accuracy score is obtained as shown in fig 12.



```
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(y_test, y_pred)
cm

array([[ 18,   8],
       [  9, 598]])

ac = accuracy_score(y_test,y_pred)
ac

0.9731437598736177
```

Fig: 12 Accuracy obtained using kNN

Decision Treealgorithm has been implemented and 97.78% accuracy score is obtained as shown in fig 13.



```
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(y_test, y_pred)
cm

array([[ 21,   5],
       [  9, 598]])

ac = accuracy_score(y_test,y_pred)
ac

0.9778830963665087
```

Fig: 13 Accuracy obtained using Decision Tree

Since accuracy obtained from Decision Tree model (97.78%) was highest, this model will be considered for our prediction model.

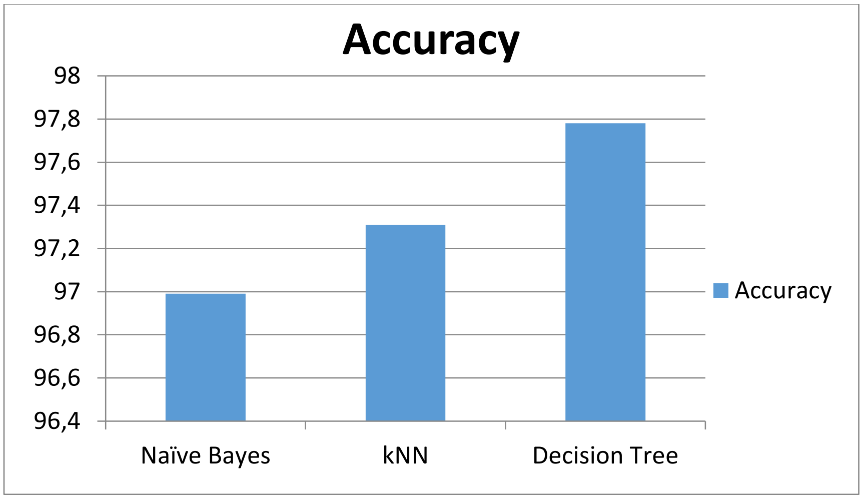


Fig: 14 Comparison of Classification algorithms based on Accuracy

5. Conclusion

Thyroid Detection using Machine Learning is a framework that aims a smart and precise way to predict thyroid disease. This paper uses Decision Tree algorithm to train the dataset and to predict thyroid disease with more accuracy. This study is concern with the motivation to develop a machine learning model to detect whether the person is normal or suffering from hypothyroidism based on their medical report. The prediction and classification of anydata depends on the data set itself and the various algorithms that are used. If anyone organizes a better data set ofreal time and applies various machine leaning algorithms such as Naive Bayes, kNN and Decision Tree then further better results may be achieved. Our objective was to give practitioner an efficient and precise way of machine learning which can be used in applications aiming to perform disease detection. Further development can be done by using image processing of ultrasonic scanning of thyroid images to predict thyroid and cancer, which cannot be recognized in blood test report. By combining both the results, thyroid disease prediction can cover all thyroid related diseases.

Reference

1. SunilaGodara, Sanjeev Kumar “Prediction of Thyroid Disease Using Machine Learning Techniques”, International Journal of Electronics Engineering (ISSN: 0973-7383) Volume 10 • Issue 2 pp. 787-793 June 2018.
2. AnkitaTyagi ,RitikaMehra , AdityaSaxena “Interactive Thyroid Disease Prediction System Using Machine Learning Technique”, 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018.
3. Chandan R , ChetanVasan , Chethan MS , Devikarani H S , “THYROID DETECTION USING MACHINE LEARNING”, International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 5, Issue 9, ISSN No. 2455-2143, Pages 173-177.
4. L. Ozyılmaz and T. Yıldırım,(2002). “Diagnosis of thyroid disease using artificial neural network methods,” 9th international conference on neural information processing 8-22 Nov. 2002.
5. V. Indumathi, S.SanthanaMegala, R.Padmapriya, M.Suganya and B.Jayanthi, “Prediction and Analysis of Plant Growth Promoting Bacteria using Machine Learning for Millet Crops”, Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 – 1833.
6. Liyong Ma , Chengkuan Ma, Yuejun Liu, and Xuguang Wang “Thyroid Diagnosis from SPECT Images Using Convolutional Neural Network with Optimization”, Hindawi Computational Intelligence and Neuroscience, 2019.
7. V. Alex, K. Vaidhya, S.)irunavukkarasu, C. Kesavadas, and G. Krishnamurthi, “Semisuperised learning using denoisingautoencoders for brain lesion detection and segmentation,” Journal of Medical Imaging, vol. 4, no. 4, article 041311, 2017.
8. Mrs. T. Pratheebha, Mrs. V. Indhumathi, Dr. S. SanthanaMegala, “An Empirical Study On Data Mining Techniques And Its Applications”, International Journal of Software & Hardware Research in Engineering (IJSHRE) ISSN-2347-4890 Volume 9 Issue 4 April 2021.
9. L.-N. Li, J.-H. Ouyang, H.-L. Chen, and D.-Y. Liu, “A computer aided diagnosis system for thyroid disease using extreme learning machine,” Journal of Medical Systems, vol. 36, no. 5, pp. 3327–3337, 2012.
10. DonthireddyShivani Reddy, OgetySaiVaishnavi, VidyaJ, K.SaiSharan, R.Subramanyam,”Diagnosis of Thyroid Gland Disorder using Machine Learning Techniques”, International Journal of Advanced Science and TechnologyVol. 29, No. 5, (2020), pp. 4752-4761
11. E. Dogantekin, A. Dogantekin, and D. Avci, “An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases,” Expert Systems with Applications, vol. 38, no. 1, pp. 146–150, 2011.
12. Z. Parry and R. Macnab, “)yroid disease and thyroid surgery,” Anaesthesia& Intensive Care Medicine, vol. 18, no. 10, pp. 488–495, 2017.
13. J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, “)yroid nodule classification in ultrasound images by finetuning deep convolutional neural network,” Journal of Digital Imaging, vol. 30, no. 4, pp. 477–486, 2017.
14. BibiAmina Begum, Prediction of thyroid Disease Using Data Mining Techniques, 5th International Conference on Advanced Computing Communication Systems (ICACCS), 2019
15. ShaikRaziaA Comparative study of machine learning algorithms on thyroid disease prediction, International Journal of Engineering & Technology, 7 (2.8) (2018) 315 – 319.