International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Constituent Depletion and Divination of Hypothyroid Prevalance using Machine Learning Classification

M. Shyamala Devi, Ankita Shil, Prakhar Katyayan, Tanmay Surana

Abstract: ***With the vast growth of technology, the world is moving towards different style of instant food habits which lead to the irregular functioning of the body organs. One such victim problem we face is the existence of hypothyroid in the body. Hypothyroid is the under active thyroid circumstance, where the thyroid gland does not produce required amount of essential hormones. The prediction of hypothyroid still remains as a challenging task due to the non availability of exact symptoms. By keeping this analysis in mind, this paper focus on prediction of hypothyroid based on the clinical parameters. The hypothyroid dataset from the UCI machine learning repository is used for predicting the existence of hypothyroid using machine learning classification algorithms. The prediction of existence of hypothyroid is carried out in four ways. Firstly, the raw data set is fitted with various classification algorithms to find the existence of hypothyroid. Secondly, the data set is tailored by the Ada Boost Regressor algorithm to extract the important features from the hypothyroid dataset. Then the extracted feature importance of the hypothyroid dataset is then fitted to the various classification algorithms. Thirdly, the hypothyroid dataset is subjected to the dimensionality reduction using principal component analysis. The PCA reduced hypothyroid dataset is then fitted with classification algorithms to predict the existence of hypothyroid. Fourth, the performance analysis is done for the raw data set, Feature importance AdaBoost hypothyroid dataset and PCA reduced hypothyroid dataset by comparing the performance metrics like precision, recall, FScore and Accuracy. This paper is implemented by python scripts in Anaconda Spyder Navigator. Experimental Result shows that the Random Forest, Naive Bayes and Logistic regression have the accuracy of 99.5 for the raw dataset, feature importance reduced dataset and the accuracy of 99.8 for the five component reduced PCA dataset.***

Index Terms: Machine Learning, Feature Extraction, PCA, MSE, MAE, R2 Score.

Revised Manuscript Received on October 30, 2019.

***** Correspondence Author

M. Shyamala Devi*, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Ankita Shil, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Prakhar Katyayan, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Tanmay Surana , III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

I. INTRODUCTION

The machine learning technology is used for finding the existence of hypothyroid using the clinical parameters. The early prediction of existence of hypothyroid still remains a challenging issue due to the lack of knowledge, neglecting the symptoms due to work pressure and the lack of exact symptoms for the prediction. The presence of hypothyroid in the body does not give any predictable symptoms to a person during initial stages of life. As the day pass on, uncared hypothyroid person may lead to several health issues like obesity, bone pain, joint pain, fertility issue, sleeping problems and heart diseases. The symptoms and existence of Hypothyroid for each person varies depending on the lack of hormones. The hypothyroid problem will tend to be slowly increasing for any person suffering from that disease.

The paper is prepared in which the survey of literatures is discussed with Section 2 go subsequently by the principal component analysis in the Section 3. Proposed work is discussed in Section 4 followed by Implementation and Performance Analysis is deliberated in Section 5 tailed by the conclusion of the paper in Section 6.

II. RELATED WORK

A. Literature Review

The neural network design models, auto immune conditions and the various condition of thyroid disease are examined. The outcome of the thyroid disease is growing drastically and provides a new path for the biological methods and to analyze the existence of thyroid disease. The various different neural network modeling are designed for the prediction of thyroid disease by using parameter estimation methods [1].

The outcome of the blood test alone cannot predict the existence of thyroid disease due to the factors like age, sex or the device used for prediction. The artificial neural networks can be applied to check the existence of thyroid disease and it can be used to find the relationship between the input attributes and the output attribute. The reliable dataset with the quantifying values are needed to predict the existence of thyroid disease. The most significant attributes of the dataset is predicted by using connection weights method and sensitivity analysis [2].

The dimensionality reduction feature extraction can be focused for the prediction of thyroid disease.

The dimensionality reduced dataset is fitted to kernelized ELM classifier for training the prediction model [3]. The concept and the usage of grafting and subdivision in the data pre-processing for the prediction of existence of thyroid disease are analyzed. The synergy effects between the two methods Re-RX with J48graft is analyzed with the classification rules [4].

The machine learning attribute collection and removal methods can be used for the detection of dependent attribute for various real time application can be understood through this article [5]–[17].

III. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is the feature extraction Dimensionality reduction method which analyzes the entire attributes of the given dataset. Then it reduces the large dataset into smaller dataset through linear transformation of attributes. The steps of Principal component analysis are shown below.

Step 1: Build the values with Covariance matrix of the given data set

Step 2: Estimate the Eigen vectors of that Covariance Matrix.

Step 3: Then the Eigen vectors with elevated and maximum Eigen values are used to restructure the data set.

Step 4: The high variance dataset features are finalized as Principal components.

IV. PROPOSED WORK

In this work, the existence of hypothyroid is predicted by using the machine learning classification algorithms. Our contribution in this paper is folded in four ways.

(i) Firstly, the raw data set is fitted with various classification algorithms to find the existence of hypothyroid and the algorithms used are as follows,

- Random Forest classifier
- Naive Bayes classifier
- Decision Tree classifier
- KNN Classifier
- Kernel SVM Classifier
- Logistic Regression Classifier

(ii) Secondly, the data set is tailored by the Ada Boost Regressor algorithm to extract the important features from the hypothyroid dataset. Then the extracted feature importance of the hypothyroid dataset is then fitted to the above mentioned classification algorithms.

(iii) Thirdly, the hypothyroid dataset is subjected to the dimensionality reduction using principal component analysis. The PCA reduced hypothyroid dataset is then fitted with classification algorithms to predict the existence of hypothyroid.

(iv) Fourth, the performance analysis is done for the raw data set, Feature importance AdaBoost hypothyroid dataset and PCA reduced hypothyroid dataset by comparing the performance metrics like Precision, recall, FScore and Accuracy

A. System Architecture

The overall design of our work is shown in Fig. 1

/

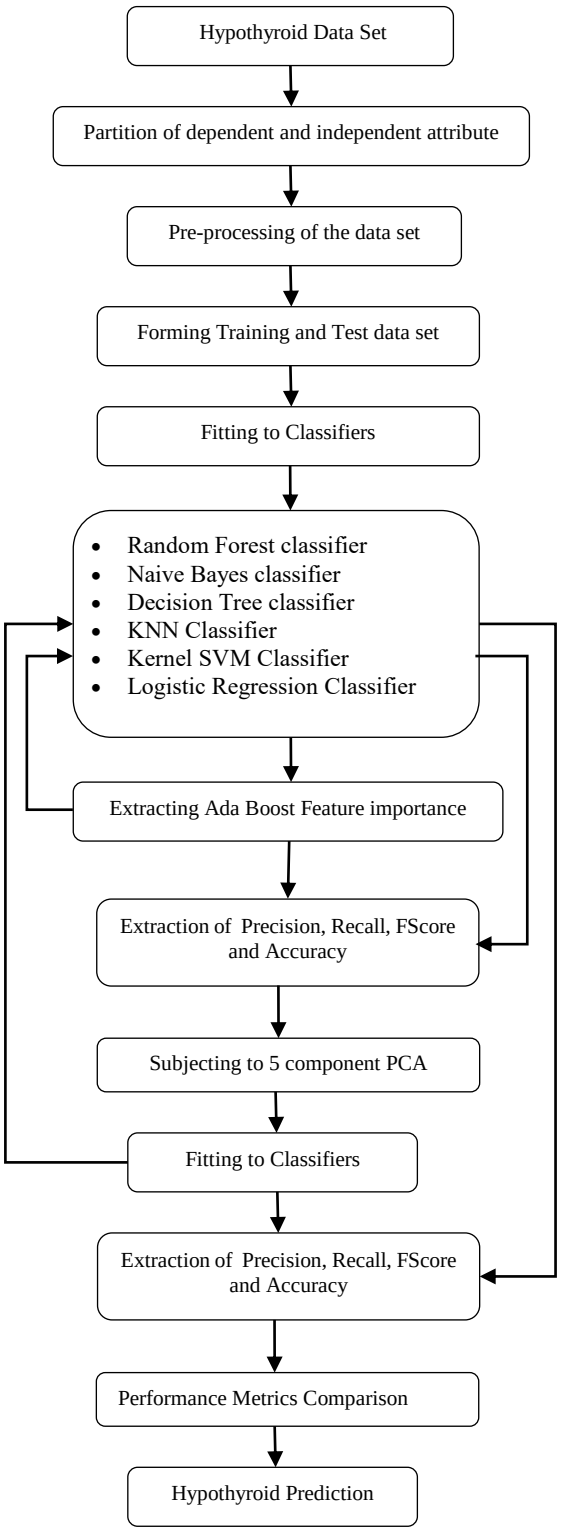

Fig. 1 System Architecture

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Hypothyroid Prediction

The Hypothyroid dataset from UCL machine learning Repository is used for implementation with 3164 instances of 23 independent attribute and 1 Hypothyroid dependent attribute.

The attribute are shown below.

1) Age
2) Sex
3) on_thyroxine
4) query_on_thyroxine
5) on_antithyroid_medication
6) thyroid_surgery
7) query_hypothyroid
8) query_hyperthyroid
9) pregnant
10) sick
11) tumor
12) lithium
13) goitre
14) TSH_measured
15) TSH- Thyroid Stimulating Hormone
16) T3_measured
17) T3- Total Triiodothyroxine
18) TT4_measured
19) TT4- Total Thyroxine
20) T4U_measured
21) T4U
22) FTI_measured
23) FTI – Free Thyroxine Index
24) TBG_measured - Thyroxine Binding Globulin - Dependent Attribute

The relationship of dataset attributes is shown in Fig 2.

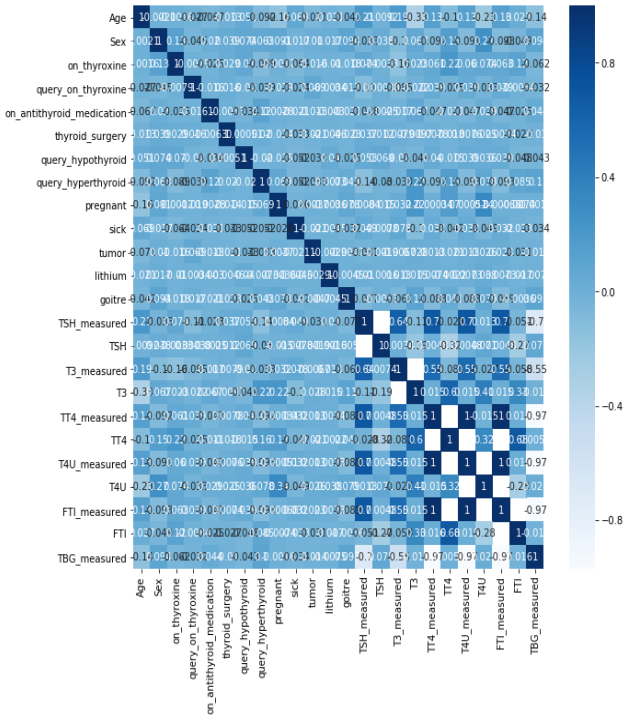

Fig. 2 Relationship of Dataset

The feature importance of the hypothyroid dataset is shown in Fig 3-Fig 5.

feature_importances - Series

Index	0
Age	0.0610272
Sex	0.027103
on_thyroxine	0.0310007
query_on_thy…	0
on_antithyro…	0
thyroid_surg…	0
query_hypoth…	0
query_hypert…	0
pregnant	0
sick	0
tumor	0
lithium	0
goitre	0
TSH_measured	0.0302475

feature_importances - Series

Index	0
sick	0
tumor	0
lithium	0
goitre	0
TSH_measured	0.0302475
TSH	0.175381
T3_measured	0.0250466
T3	0.00976049
TT4_measured	0.000525936
TT4	0.0801387
T4U_measured	0.277919
T4U	0.148022
FTI_measured	0.0316351
FTI	0.102193

Fig. 3 Feature importance of the hypothyroid dataset

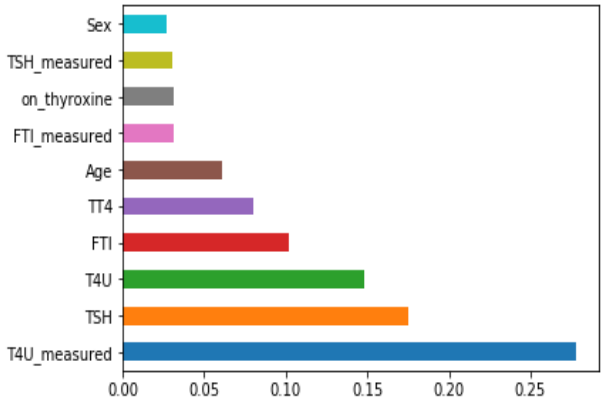

Fig. 4 Extraction of Feature importance

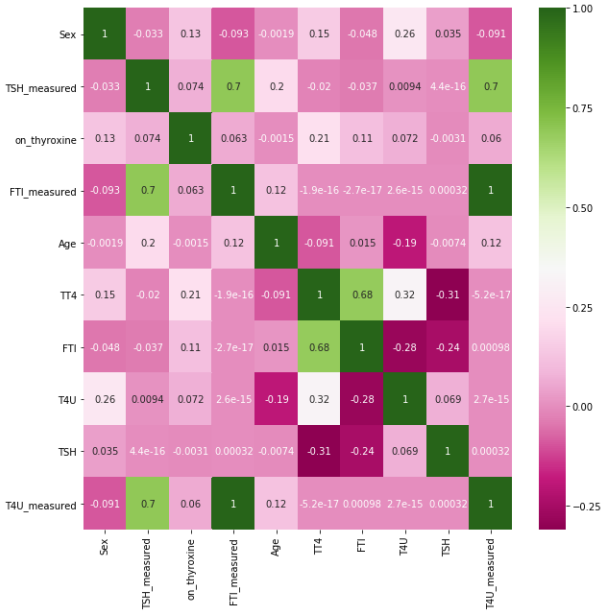

Fig.5 Correlation of Feature importance

The confusion matrix obtained for the classifiers of the raw dataset is shown n fig 6.

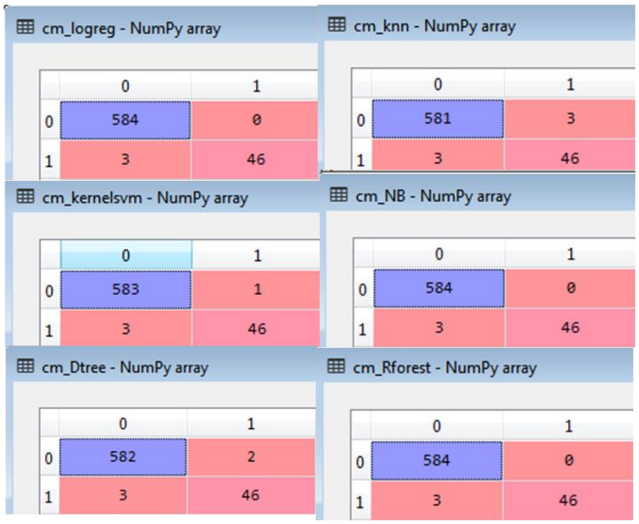

Fig. 6 Raw dataset Confusion matrix

The confusion matrix obtained for the classifiers of the Ababoost Feature importance is shown n fig 7.

Fig. 7 Ababoost Feature importance Confusion matrix

The relationship of the attribute variance and the PCA component is shown in Fig 8.

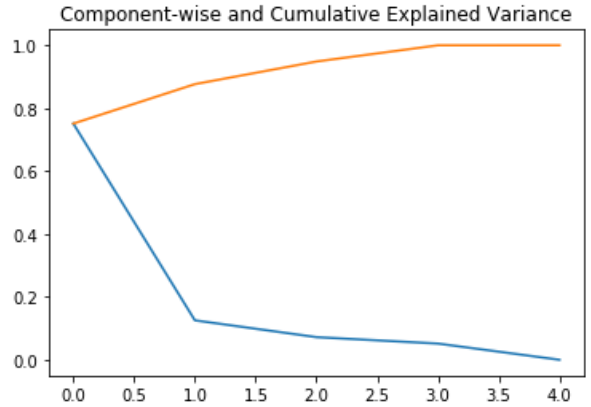

Fig. 8 Attribute variance VS PCA component

The number of components that can be applied using PCA is analyzed with the elbow method and is shown in Fig 9.

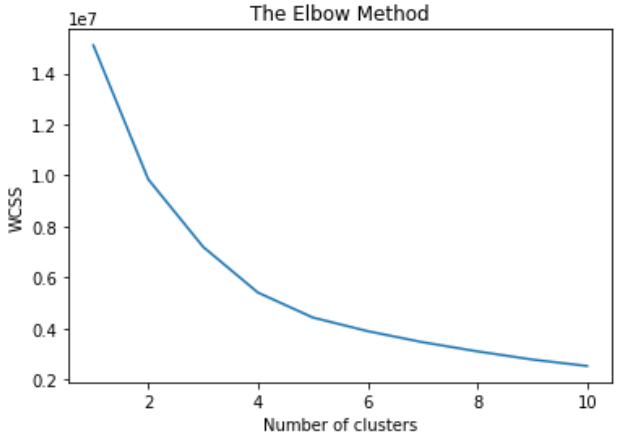

Fig. 9 Identification of PCA components

The clustering of thyroid patients in the dataset is shown in Fig 10.

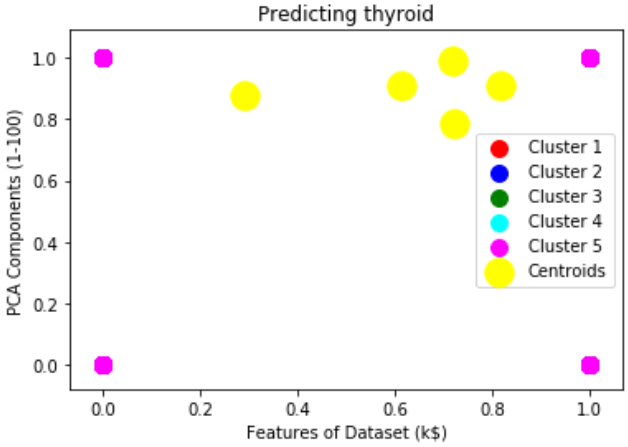

Fig. 9 Clustering of thyroid patients.

The confusion matrix obtained for the classifiers of the five component PCA is shown n fig 11.

Fig.11 Five Component PCA Confusion matrix

B. Analysis of Parameters Efficiency

The raw data set is fitted with various classification algorithms to predict the following metrics shown in Table 1.

Table. 1. Applying Raw dataset to classifiers

Classifiers	Applying Raw Data Set to Classifiers			
	Precision	Recall	FScore	Accuracy
RandomForest	0.99	1.00	1.00	99.5
DecisionTree	0.99	1.00	1.00	99.2
NaiveBayes	0.99	1.00	1.00	99.5
KNN	0.99	0.99	0.99	99.1
KerneSVM	0.98	1.00	1.00	99.4
Logistic	0.99	1.00	1.00	99.5

The feature importance adaboost data set is fitted with various classification algorithms to predict the following metrics shown in Table 2.

Table. 2. Applying feature importance dataset to classifiers

Classifiers	Applying adaboost dataset to Classifiers			
	Precision	Recall	FScore	Accuracy
RandomForest	0.99	1.00	1.00	99.5
DecisionTree	0.99	0.99	0.99	98.9
NaiveBayes	0.99	1.00	1.00	99.5
KNN	0.99	0.99	0.99	99.1
KerneSVM	0.99	1.00	1.00	99.5
Logistic	0.99	1.00	1.00	99.5

The 5 component PCA reduced dataset is fitted with various classification algorithms to predict the following metrics shown in Table 3.

Table. 3. Applying 5 component PCA dataset to classifiers

Classifiers	Applying 5 component PCA dataset to Classifiers			
	Precision	Recall	FScore	Accuracy
RandomForest	1.00	1.00	1.00	99.8
DecisionTree	1.00	0.99	1.00	99.2
NaiveBayes	1.00	0.98	0.99	98.4
KNN	1.00	0.99	1.00	99.2
KerneSVM	1.00	1.00	1.00	99.8
Logistic	1.00	1.00	1.00	99.8

The performance analysis of the efficiency metrics is shown in the Fig 12 – Fig 15.

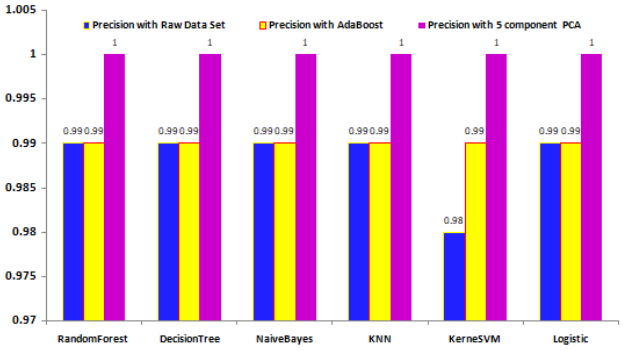

Fig. 12 Analysis of Precision

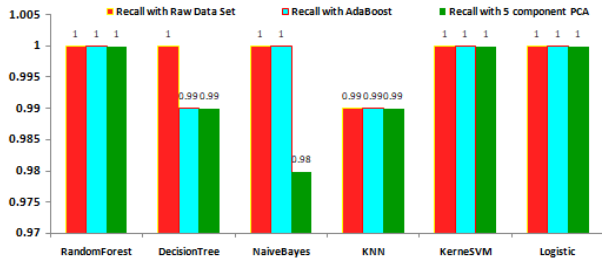

Fig. 13 Analysis of Recall

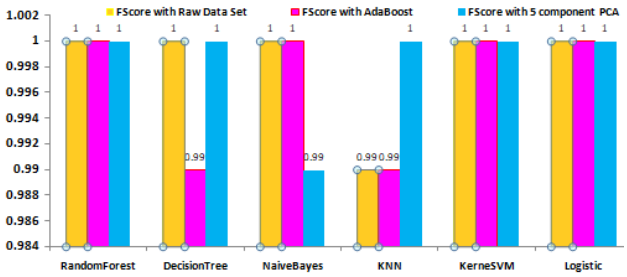

Fig. 14 Analysis of FScore

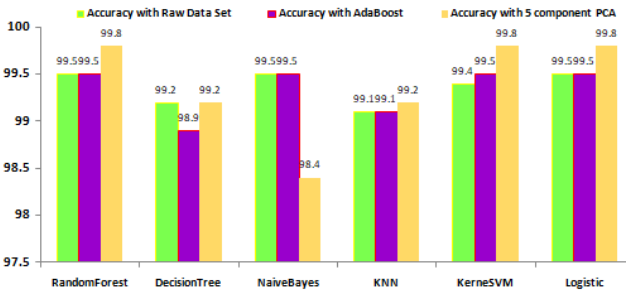

Fig. 15 Analysis of Accuracy

VI. CONCLUSION

This paper analyzes the performance of hypothyroid identification for subjecting the various classification algorithms to the raw dataset, feature importance reduced dataset, feature extraction dimensionality reduced Principal component analysis dataset. The performance analysis is done with the metrics such as Accuracy, Recall, Precision and FScore. Experimental Result shows that the Random Forest, Naive Bayes and Logistic regression have the accuracy of 99.5 for the raw dataset, feature importance reduced dataset and the accuracy of 99.8 for the five component reduced PCA dataset.

REFERENCES

1. Shaik Razia and M. R. Narasinga Rao, "Machine Learning Techniques for Thyroid Disease Diagnosis - A Review" ",Indian Journal of Science and Technology., Vol. 9, no. 28, July 2016.
2. Martyna Michałowska, T. Walczak, Jakub Krzyszt and Monika Grygorowicz, "Assessment of Clinical Variables Importance with the Use of Neural Networks by the Example of Thyroid Blood Test Parameters", Innovations in Biomedical Engineering, August 2019.
3. Chao Ma, Jian Guan, Wenyong Zhao and Chaolun Wang, "An Efficient Diagnosis System for Thyroid Disease Based on Enhanced Kernelized Extreme Learning Machine Approach", Cognitive Computing, June 2018

4. Yoichi Hayashi, "Synergy effects between grafting and subdivision in Re-RX with J48graft for the diagnosis of thyroid disease", Knowledge-Based Systems, June 2017.
5. R. Suguna, M. Shyamala Devi, and Rincy Merlin Mathew, " Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.
6. Suguna Ramadass, and Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers", Springer's book series "Learning and Analytics in Intelligent Systems, Springer, , LAIS vol. 3, pp. 613-620, 2019.
7. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, and Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739.
8. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 6198-6203.
9. M. Shyamala Devi, Rincy Merlin Mathew, and R. Suguna,"Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
10. Shyamala Devi Munisamy, and Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification", Springer's book series "Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 604-612, 2019.
11. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019.pp. 4800-4807.
12. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Sairam Kondapalli, "Recognition of forest Fire Spruce Type Tagging using Machine Learning Classification", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 30 September 2019.
13. M. Shyamala Devi, Usha Vudatha, Sukriti Mukherjee, Bhavya Reddy Donthiri, S B Adhiyan, Nallareddy Jishnu, " Linear Attribute Projection and Performance Assessment for Signifying the Absenteeism at Work using Machine Learning", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 30 September 2019.
14. M. Shyamala Devi, Mothe Sunil Goud, G. Sai Teja, MallyPally Sai Bharath, "Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.11, 30 September 2019
15. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, " Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 952 – 956, 30 August 2019
16. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 957 – 961, 30 August 2019.
17. Rincy Merlin Mathew, R. Suguna, M. Shyamala Devi, "Composite Model Fabrication of Classification with Transformed Target Regressor for Customer Segmentation using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 962 – 966, 30 August 2019.