International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Feature Scaled Element Balancing with Random Boosting for Heart Disease Prediction using Machine Learning

M. Shyamala Devi, Shermin Shamsudheen, Rincy Merlin Mathew

Abstract*: In the current scenario, the researchers are focusing towards health care project for the prediction of the disease and its type. In addition to the prediction, there exists a need to find the influencing parameter that directly related to the disease prediction. The analysis of the parameters needed to the prediction of the disease still remains a challenging issue. With this view, we focus on predicting the heart disease by applying the dataset with boosting the parameters of the dataset. The heart disease data set extracted from UCI Machine Learning Repository is used for implementation. The anaconda Navigator IDE along with Spyder is used for implementing the Python code. Our contribution is folded is folded in three ways. First, the data preprocessing is done and the attribute relationship is identified by the correlation values. Second, the data set is fitted to random boost regressor and the important features are identified. Third, the dataset is feature scaled reduced and then fitted to random forest classifier, decision tree classifier, Naïve bayes classifier, logistic regression classifier, kernel support vector machine and KNN classifier. Fourth, the dataset is reduced with principal component analysis with five components and then fitted to the above mentioned classifiers. Fifth, the performance of the classifiers is analyzed with the metrics like accuracy, recall, fscore and precision. Experimental results shows that, the Naïve bayes classifier is more effective with the precision, Recall and Fscore of 0.89 without random boost, 0.88 with random boosting and 0.90 with principal component analysis. Experimental results show, the Naïve bayes classifier is more effective with the accuracy of 89% without random boost, 90% with random boosting and 91% with principal component analysis.

***Index Terms*: Machine Learning, Accuracy, Performance Metrics, Regressor and Feature scaling.**

I. INTRODUCTION

The heart disease is one of the major chronic diseases in the today's technological advancement. There are various heart diseases that may be the reason for the arrest of the heart functioning. The heart disease may also make the brain and kidney to stop functioning. The strengthening of the heart blood vessels may lead to knock heart disease which is initiated by the blood pressure of the body. The pressure and the hemoglobin is the major reason for the immunization of the body. So it is the responsibility of the medical researchers to find the parameter dependency in finding the existence of the heart disease. The machine learning technology can be used for predicting the type of the heart disease and the existence of the particular class of heart disease in advance. This advance prediction helps the patients to prepare themselves in protecting themselves against the disease. The various heart diseases are given awareness towards the people in shake of enhancing the wealth of the nation. World Health Organization is finding many solutions towards creating awareness to the people in forming the healthiest nation.

The paper is organized in such a way that Section 2 deals with the related works. Section 3 discuss about the proposed work followed by the implementation and Performance Analysis in Section 4. The paper is concluded with Section 5.

II. RELATED WORK

A. Literature Review

The heart disease prediction is done based on the concept of dissimilar classification algorithms using various classifiers. The classifiers used for the prediction of heart disease are Decision Stump, Random Forest and LMT tree algorithm. The performance comparison of the dissimilar classification algorithms are done based on the accuracy using weka tool. The experimental results were combined for all the dissimilar classification algorithms and they are used to classify the type and class of heart disease [1].

The data mining algorithms can be used to forecast the existence of the heart disease using the dissimilar data mining classifier tools. The other parts of the body stops functioning if anything happens to the functionality of the heart. The heart disease is the major reason for the death of any patients. So the prediction of the heart disease at the earliest stage may help the individual to extend their life and to live happily [2].

The heart disease is the resultant of tobacco chewing and inhaling of drugs by the individual. But the exact cause of the heart disease cannot be predicted appropriately as the heart disease may come due to variety of food habits, life style and lack of physical activity. It stills remains the challenge in the medical records to find the exact occurrence of the heart disease.

Manuscript published on January 30, 2020.
* 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.
Shermin Shamsudheen, Lecturer, Department of Computer Science, College of Computer Science & Information Systems, Jazan University, Saudi Arabia.
Rincy Merlin Mathew, Lecturer, Department of Computer Science, College of Science and Arts, Khamis Mushayt, King Khalid university, Abha, Asir, Saudi Arabia.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

They are various types of mathematical data mining tools towards predicting the heart disease in advance [3]. The machine learning feature selection and feature extraction methods can be used for the prediction of any factor in different application can be learnt from this article [4] – [28].

III. PROPOSED WORK

In our proposed work, the heart disease dataset is boosted with Random boost regressor and then feature reduced with prinicipal component analysis to predict the class of heart disease. Our implementation in this paper is folded in five ways.

(i) First, the data preprocessing is done and the attribute relationship is identified by the correlation values.
(ii) Second, the data set is fitted to random boost regressor and the important features are identified.
(iii) Third, the dataset is feature scaled reduced and then fitted to random forest classifier, decision tree classifier, Naive bayes classifier, logistic regression classifier, kernel support vector machine and KNN classifier.
(iv) Fourth, the dataset is reduced with principal component analysis with five components and then fitted to the above mentioned classifiers.
(v) Fifth, the performance of the classifiers is analyzed with the metrics like accuracy, recall, fscore and precision.

A. System Architecture

The overall block diagram of the proposed work is shown in Fig. 1.

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Target Transformation Heart Disease Dataset

The Heart Disease dataset extracted from UCL ML Repository is used for implementation with 13 independent attribute and 1 diagnosis dependent attribute. The dataset consists of 779 individual's data. The attribute are shown below. The classes of the dependent attribute are shown in Fig. 2.

1.	Age
2.	Sex
3.	Chest-pain type
4.	Resting Blood Pressure
5.	Serum Cholestrol
6.	Fasting Blood Sugar
7.	Resting ECG
8.	Max heart rate
9.	Angina
10.	ST depression
11.	Peak exercise ST segment
12.	Number of major vessels
13.	Thal
14.	Diagnosis of heart disease - Dependent Attribute

Chest Pain Format	Desription
1	Typical Angina
2	Atypical Angina
3	Non - Anginal Pain
4	Asymptotic

Fig. 2. Pain Type Class Levels

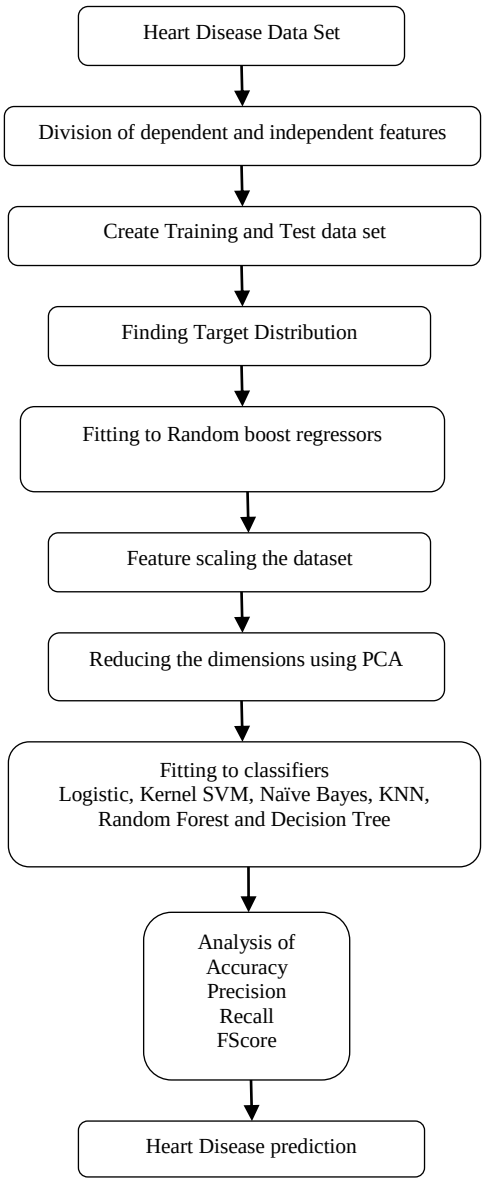


Fig. 1 Work Flow of the Model

The division of the target heart disease class type variable is shown in Fig. 3.

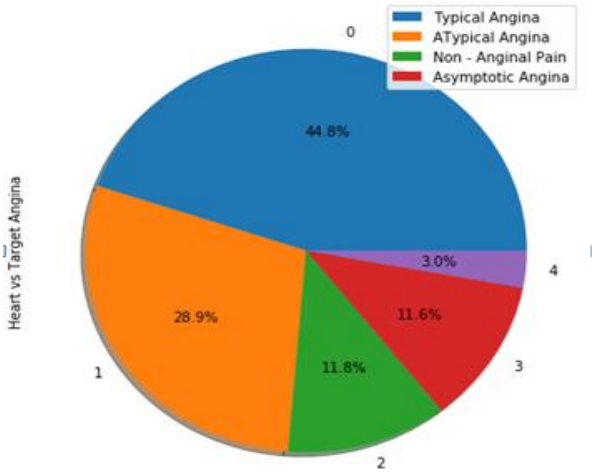


Fig. 3. Class Division

The dependency of each parameters is shown in Fig. 4.

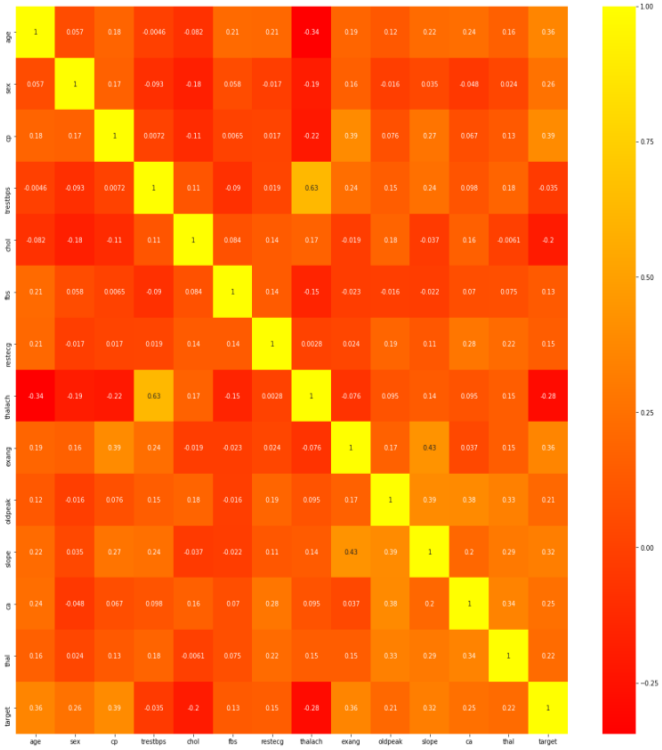


Fig. 4 Correlation Matrix of Heart Disease dataset

The raw dataset is applied to the random boost regressor to show the relationship of the feature important attributes of the dataset and is shown in the Fig. 5.

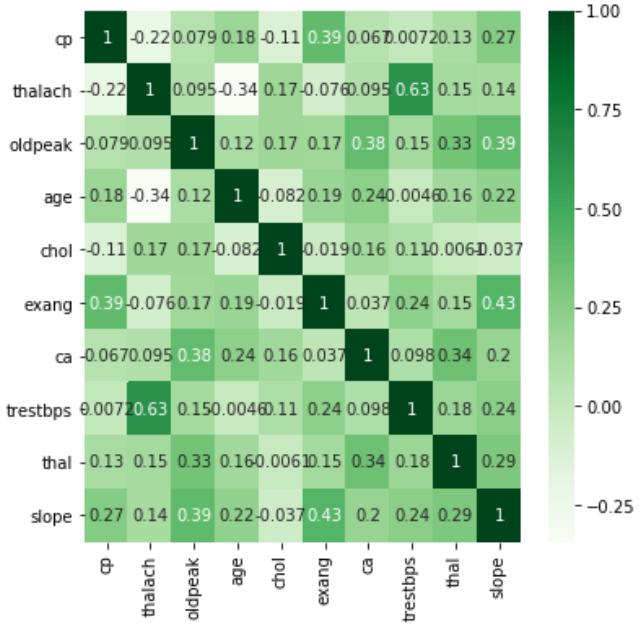
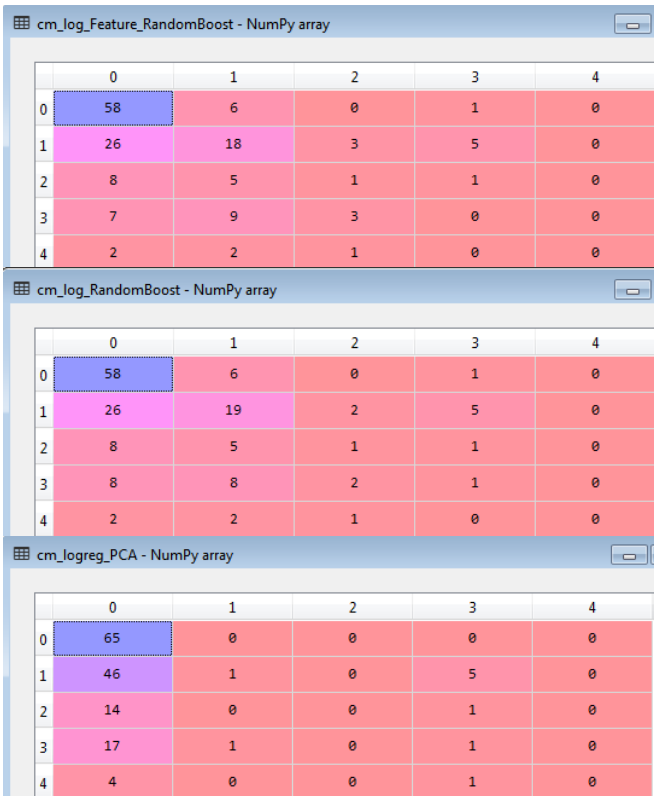


Fig. 5. Features of Random Boost Regressor

The raw dataset, Dimensionality reduced Principal component analysis dataset and Features of the Random Boost Regressor dataset are fitted to random forest classifier, decision tree classifier, Naive bayes classifier, logistic regression classifier, kernel support vector machine and KNN classifier and the performance indicator is shown in Fig. 6 – Fig 12.



cm_log_Feature_RandomBoost - NumPy array

	0	1	2	3	4
0	58	6	0	1	0
1	26	18	3	5	0
2	8	5	1	1	0
3	7	9	3	0	0
4	2	2	1	0	0

cm_log_RandomBoost - NumPy array

	0	1	2	3	4
0	58	6	0	1	0
1	26	19	2	5	0
2	8	5	1	1	0
3	8	8	2	1	0
4	2	2	1	0	0

cm_logreg_PCA - NumPy array

	0	1	2	3	4
0	65	0	0	0	0
1	46	1	0	5	0
2	14	0	0	1	0
3	17	1	0	1	0
4	4	0	0	1	0

Fig. 6. Logistic Regression Classifier

cm_log_KNN_RandomBoost - NumPy array

	0	1	2	3	4
0	50	14	1	0	0
1	31	16	2	3	0
2	9	5	0	1	0
3	7	11	0	1	0
4	3	1	0	1	0

cm_log_KNNFeatureRandomBoost - NumPy array

	0	1	2	3	4
0	58	6	0	0	1
1	25	20	5	2	0
2	8	4	3	0	0
3	3	9	5	2	0
4	1	1	3	0	0

cm_knn_PCA - NumPy array

	0	1	2	3	4
0	53	10	2	0	0
1	37	8	5	2	0
2	12	2	1	0	0
3	11	6	1	1	0
4	5	0	0	0	0

Fig. 7 KNN Classifier

cm_KSVM_RandomBoost - NumPy array

	0	1	2	3	4
0	65	0	0	0	0
1	51	1	0	0	0
2	14	0	1	0	0
3	19	0	0	0	0
4	5	0	0	0	0

cm_FeatureKSVM_randomBoost - NumPy array

	0	1	2	3	4
0	57	7	0	1	0
1	23	25	2	2	0
2	6	8	1	0	0
3	2	12	4	1	0
4	1	2	2	0	0

cm_kernelsvm_PCA - NumPy array

	0	1	2	3	4
0	59	6	0	0	0
1	41	7	2	2	0
2	11	2	2	0	0
3	13	5	1	0	0
4	5	0	0	0	0

Fig. 8. Kernel SVM Classifier

cm_NaiveBayes_RandomBoost - NumPy array

	0	1	2	3	4
0	52	8	2	0	3
1	15	26	3	6	2
2	7	5	2	1	0
3	2	9	5	1	2
4	0	2	2	1	0

cm_FeatureNaiveBayes_RandomBoost - NumPy array

	0	1	2	3	4
0	52	8	2	0	3
1	15	26	3	6	2
2	7	5	2	1	0
3	2	9	5	1	2
4	0	2	2	1	0

cm_NB_PCA - NumPy array

	0	1	2	3	4
0	59	6	0	0	0
1	30	16	0	6	0
2	7	7	0	1	0
3	11	7	0	1	0
4	4	0	0	1	0

Fig. 9.Naive Bayes Classifier

cm_DecisionTree_RandomBoost - NumPy array

	0	1	2	3	4
0	44	15	3	1	2
1	22	17	7	6	0
2	3	3	7	2	0
3	6	3	5	3	2
4	1	2	1	1	0

cm_FeatureDecisionTree_RandomBoost - NumPy array

	0	1	2	3	4
0	44	15	3	1	2
1	22	17	7	6	0
2	3	3	7	2	0
3	6	3	5	3	2
4	1	2	1	1	0

cm_Dtree_PCA - NumPy array

	0	1	2	3	4
0	27	22	8	7	1
1	20	16	5	10	1
2	6	2	4	3	0
3	9	7	3	0	0
4	4	1	0	0	0

Fig. 10.Decision Tree Classifier

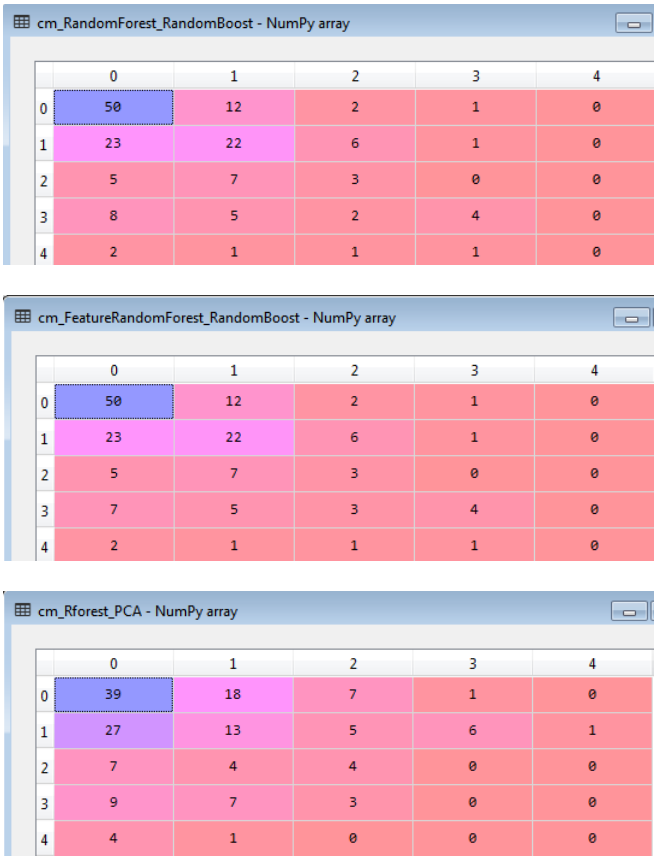


Fig. 11.Random Forest Classifier

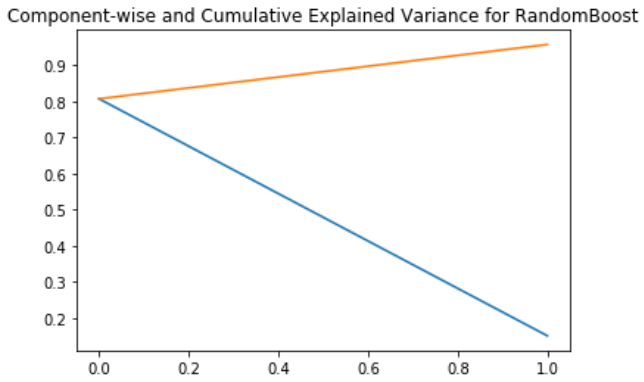


Fig. 12. PCA Graph indicator

The performance of the classifiers is analyzed with the metrics like accuracy, recall, fscore and precision and is shown in Table. 1 – Table. 4.

Table. 1 Precision Comparison for Random Boost

Classifiers	Precision		
	Random Boost	Feature Scaled Random Boost	PCA
Log Reg	0.77	0.76	0.79
KNN	0.81	0.82	0.82
Kernel SVM	0.79	0.78	0.79
Naïve Bayes	0.89	0.88	0.90
Decision Tree	0.86	0.85	0.84
Random Forest	0.88	0.87	0.87

Table. 2 Recall Comparison for Random Boost

Classifiers	Recall		
	Random Boost	Feature Scaled Random Boost	PCA
Log Reg	0.76	0.76	0.79
KNN	0.80	0.82	0.82
Kernel SVM	0.79	0.78	0.79
Naïve Bayes	0.89	0.88	0.90
Decision Tree	0.87	0.84	0.84
Random Forest	0.87	0.86	0.86

Table. 3 FScore Comparison for Random Boost

Classifiers	FScore		
	Random Boost	Feature Scaled Random Boost	PCA
Log Reg	0.76	0.75	0.79
KNN	0.80	0.81	0.82
Kernel SVM	0.78	0.77	0.79
Naïve Bayes	0.89	0.88	0.90
Decision Tree	0.85	0.84	0.83
Random Forest	0.86	0.86	0.86

Table. 4 Accuracy Comparison for Random Boost

Classifiers	Accuracy (%)		
	Random Boost	Feature Scaled Random Boost	PCA
Log Reg	76	77	78
KNN	81	82	80
Kernel SVM	79	80	79
Naïve Bayes	89	90	91
Decision Tree	86	85	84
Random Forest	85	86	84

V. CONCLUSION

This paper endeavour to predit the type of heart disease by considering the features that are essentially needed for prediction of heart disease. The dependency of the parameters is reviewed with the correlation matrix and correlation values of each of the dataset attributes. Experimental results show that, the Naïve bayes classifier is more effective with the precision, Recall and Fscore of 0.89 without random boost, 0.88 with random boosting and 0.90 with principal component analysis. Experimental results show, the Naïve bayes classifier is more effective with the accuracy of 89% without random boost, 90% with random boosting and 91% with principal component analysis.

REFERENCES

1. S. Vijiyarani et. al., “An Efficient Classification Tree Technique for Heart Disease Prediction”,International Conference on Research Trends in Computer Technologies (ICRTCT - 2013), Proceedings of the International Journal of Computer Applications (0975 – 8887), 2013, pp. 6-9.
2. KRISHNAIAH, G. NARASIMHA AND N. SUBHASH CHANDRA, “HEART DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUES AND INTELLIGENT FUZZY APPROACH: A REVIEW” INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS, 2016.
3. VikasChaurasia and Saurabh Pal, “Early Prediction of Heart disease using Data mining Techniques”, Caribbean journal of Science and Technology, 2013.

4. M. Shyamala Devi, Shakila Basheer, Rincy Merlin Mathew, "Exploration of Multiple Linear Regression with Ensembling Schemes for Roof Fall Assessment using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019.
5. Shakila Basheer, Rincy Merlin Mathew, M. Shyamala Devi, "Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 127-133.
6. Rincy Merlin Mathew, M. Shyamala Devi, Shakila Basheer," Exploration of Neighbor Kernels and Feature Estimators for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 597-605.
7. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Nariboyena Vijaya Sai Ram, "Backward Eliminated Formulation of Fire Area Coverage using Machine Learning Regression", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp.1565-1569
8. M. Shyamala Devi, Ankita Shil, Prakhar Katyayan, Tanmay Surana, "Constituent Depletion and Divination of Hypothyroid Prevalance using Machine Learning Classification", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 1607-1612
9. 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, pp. 4309 – 4313, 16 September 2019.
10. 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, pp. 1262 – 1267, 16 September 2019.
11. 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, pp. 604 – 609, 30 September 2019.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, 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. DOI: 10.1080/09720510.2019.1609729ISSN: 0972-0510 (Print), 2169-0014 (Online).
18. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, 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. DOI: 10.1080/09720510.2019.1609729ISSN: 0972-0510 (Print), 2169-0014 (Online).
19. Shyamala Devi Munisamy, Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification", Learning and Analytics in Intelligent Systems, LAIS, Springer, vol. 3, pp. 604-612, June 2019.
20. Suguna Ramadass, 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 entitled "Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 613-620, June 2019.
21. M. Shyamala Devi, Rincy Merlin Mathew, 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.
22. R. Suguna, M. Shyamala Devi, 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.
23. Kota Venkateswara Rao, M. Shyamala Devi," Interrogation of Sentiment Perusing with Hash Counting Vectorizer and Term Inverse Frequency Transformer using Machine Learning Classification", International Journal of Recent Technology and Engineering, Volume-8 Issue-4, , 30 November 2019, pp. 3895 - 3901.
24. M. Shyamala Devi, Jyotikinkar Saharia, Shubham Kumar, Aayushi Chansoriya, Prashant Yadav, "Machine Learning Based Suspicion of Customer Detention in Banking with Diverse Solver Neighbors and Kernels", International Journal of Recent Technology and Engineering, Volume-8 Issue-4, , 30 November 2019, pp. 3244 - 3249.
25. M. Shyamala Devi, G. Bhargava Krishna, K.Sowmya, T. Sabari Pavan," Postulation of Customer Retention in Banking Sector using Machine Learning and Principal Component", International Journal of Recent Technology and Engineering, Volume-8 Issue-4, , 30 November 2019, pp. 3178- 3182.
26. M. Shyamala Devi, Nariboyena Vijaya Sai Ram, Aravapalli Sai Vamshi, Basyam Bharath, Mallangi Surya Prakash Reddy, "Swindling Shonky Anatomization of Credit Card Transactions using Machine Learning" International Journal of Recent Technology and Engineering, Volume-8 Issue-4, , 30 November 2019, pp. 1477- 1483.
27. M. Shyamala Devi, Kamma Rahul, Manubolu Satheesh, Koruprolu Rajasekhar, Pittala Ganesh Kumar," Count Vectorized Spam and Ham Discernment of Short Message Service using Machine Learning Classification", International Journal of Recent Technology and Engineering, Volume-8 Issue-4, 30 November 2019, pp. 557- 561
28. Kota Venkateswara Rao, M. Shyamala Devi," Disquisition of Sentiment Inquiry with Hashing and Counting Vectorizer using Machine Learning Classification", International Journal of Innovative Technology and Exploring Engineering, vol. 9, no.1, 10 November 2019, pp. 737-743.