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

Attribute Balanced Leveling with Ada Boost Regressor for Predicting Heart Disease using Machine Learning

Shermin Shamsudheen, Rincy Merlin Mathew, M. Shyamala Devi

Abstract*: The technological advancement can help the entire application field to predict the damage and to forecast the future target of the object. The wealth of the world is in the health of the people. So the technology must support the technologists in predicting the disease in advance. The machine learning is the emerging field which is used to forecast the existence of the heart disease through the values of the clinical parameters. With this view, we focus on predicting the customer churn for the banking application. This paper uses the customer churn bank modeling data set extracted from UCI Machine Learning Repository. 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 is processed to find the relationship between the elements of the dataset. Second, the data set is applied for Ada Boost regressors and the important elements are identified. Third, the dataset is applied to feature scaling and then fitted to kernel support vector machine, logistic regression classifier, Naive bayes classifier, random forest classifier, decision tree classifier and KNN classifier. Fourth, the dataset is dimensionality reduced with principal component analysis with five components and then applied to the previously mentioned classifiers. Fifth, the performance of the classifiers is analyzed with the indication metrics like precision, accuracy, recall and Fscore. The implementation is carried out with python code using Anaconda Navigator. Experimental results show that, the Naïve bayes classifier is more effective with the precision of 0.90 for dataset with random boost, feature scaled and PCA. Experimental results show that, the Naïve bayes classifier is more effective with the recall of 0.91 for dataset with random boost, feature scaled and PCA. Experimental results show that, the Naïve bayes classifier is more effective with the Fscore of 0.92 for dataset with random boost, feature scaled and PCA. Experimental results show, the Naïve bayes classifier is more effective with the accuracy of 91% without random boost, 93% with random boosting and 92% with principal component analysis.

Index Terms*: Machine Learning, Churn, Classification, accuracy, precision, recall and f-score.

I. INTRODUCTION

The machine learning technology is used for various applications for the prediction of the existence of disease in the human body in advance. All the forecasting applications can use the machine learning algorithms for the prediction and the analysis of various significance changes that can occur in near future. The medical history and the events of the past patients can be used to train the machine for forecasting the nature of the disease for the new patients. This type of training the machine can be used to predict exactly for the testing data of new patients.

The paper is organized in such a way that Section 2 deals with the related works. Section 3 **discuss a**bout 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 rule of relationship over the parameters of the heart disease dataset is done to predict the type of heart disease. The risk factors like narrowed artery and heart perfusion were also accessed to refine the dataset. All the dataset features were assigned to the single rule. This rule will be distinctly identify and distinguish the attributes of the heart disease dataset. The numbers of features that are originated from the rule are organized by the medical history of the patients and the clustered rules [1].

The data mining technology was one of the technologies which provide medical data mining thereby creating the medical awareness towards the public. Recently, the medical field is producing large amounts of data. However all the medical data are just stored and it is kept unused. With the advent of the technology, the early disease prediction can be done using machine learning and data mining. The variations in the clinical parameters of the patients can be taken into consideration for predicting the availability of such heart disease [2].

Heart is the essential organ for the human being to exist. So, predicting the heart disease will be significant research in the field of medical health care. Data analytics can be used to interpret the various attributes that are related to the heart disease. The dataset can be preserved to find the existence of future diseases. The heart disease can be well predicted by the artificial neural network [3].

Revised Manuscript Received on January 15, 2020.

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.

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

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 is processed to find the relationship between the elements of the dataset.

(ii) Second, the data set is applied for Ada Boost regressors and the important elements are identified and displayed as a chart.

(iii) Third, the dataset is applied to feature scaling and then fitted to kernel support vector machine classifier, logistic regression classifier, Naive bayes classifier, random forest classifier, decision tree classifier and KNN classifier.

(iv) Fourth, the dataset is dimensionality reduced with principal component analysis with five components and then applied to the previously mentioned classifiers.

(v) Fifth, the performance of the classifiers is analyzed with the indication metrics like precision, accuracy, recall and Fscore.

A. System Architecture

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

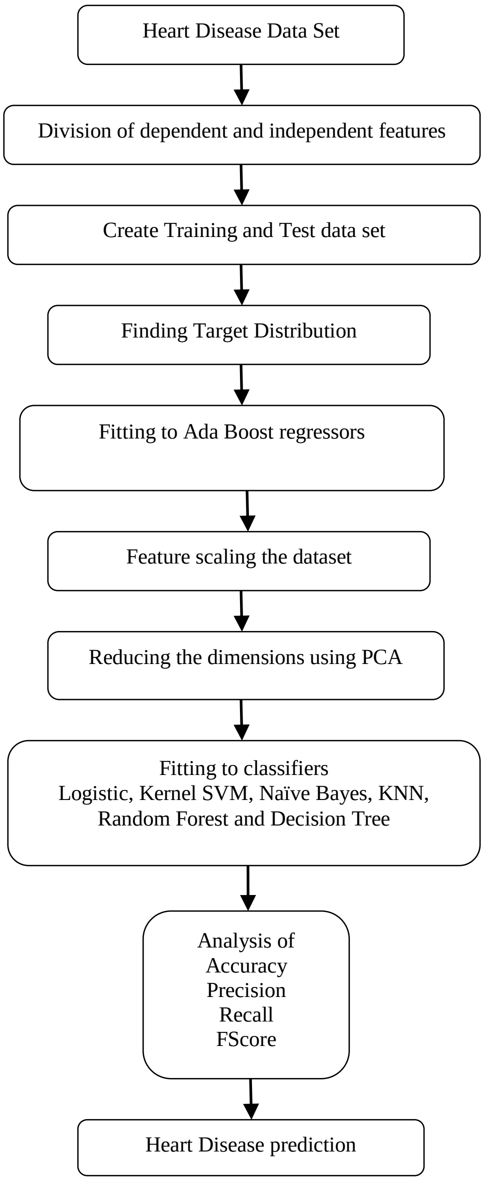


Fig. 1 Work Flow of the Model

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

The existence and amount of target heart disease class type variable is shown in Fig. 3.

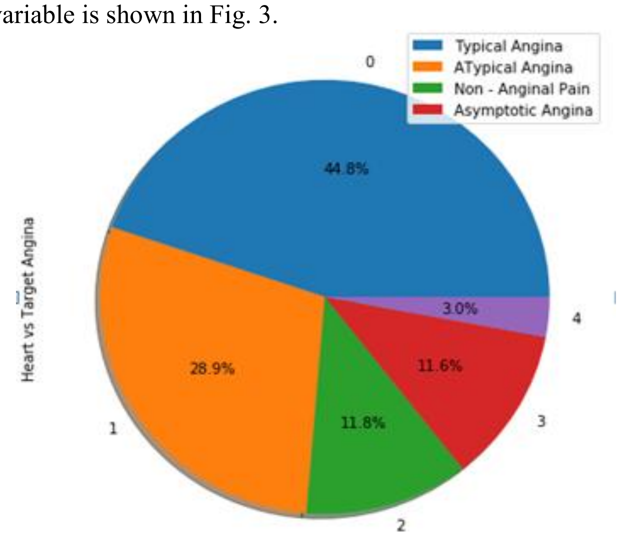


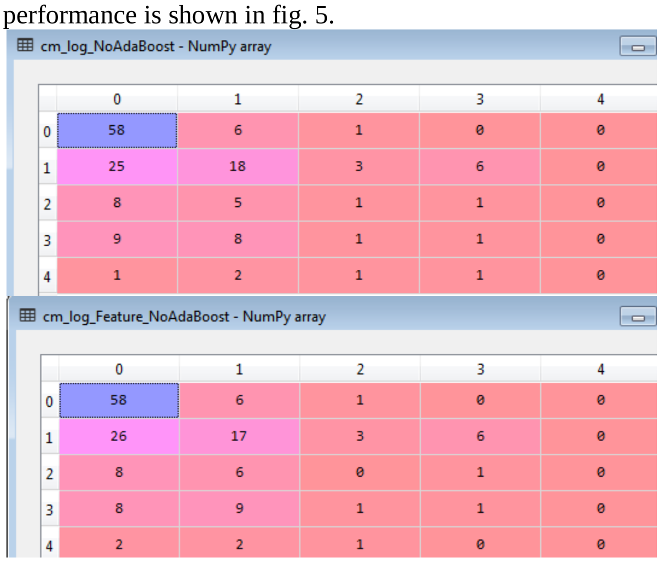
Fig. 3. Target Class Distribution

The expansion of the dataset attributes shown in Fig. 4.

S.No	Attribute	Description
1.	Age	Age of the individual.
2.	Sex	Gender of the individual with the following format: 1 = male 0 = female.
3.	Chest-pain type	Type of chest-pain as in Table. 2.
4.	Resting Blood Pressure	Resting blood pressure value in mmHg (unit)
5.	Serum Cholestrol	serum cholestrol in mg/dl (unit)
6.	Fasting Blood Sugar	Fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false)
7.	Resting ECG	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
8.	Max heart rate	Max Heart Rate
9.	Angina	1 = yes 0 = no
10.	ST depression	Value (integer or float).
11.	Peak exercise ST segment	1 = Upsloping 2 = Flat 3 = Downsloping
12.	Number of major vessels	Values (0-3) colored by flourosopy
13.	Thal	displays the Thalassemia : 3 = Normal 6 = Fixed Defect 7 = Reversable Defect
14.	Diagnosis of heart disease	0 = Absence 1,2,3,4 = Present.

Fig. 4. Heart Disease Schema Design

The raw dataset is applied to the classifier and the performance is shown in fig. 5.



cm_log_NoAdaBoost - NumPy array

	0	1	2	3	4
0	58	6	1	0	0
1	25	18	3	6	0
2	8	5	1	1	0
3	9	8	1	1	0
4	1	2	1	1	0

cm_log_Feature_NoAdaBoost - NumPy array

	0	1	2	3	4
0	58	6	1	0	0
1	26	17	3	6	0
2	8	6	0	1	0
3	8	9	1	1	0
4	2	2	1	0	0

Fig. 5. Classifier output with no ada boost

The data set is applied for Ada Boost regressors and the important elements are is shown in the Fig. 6.

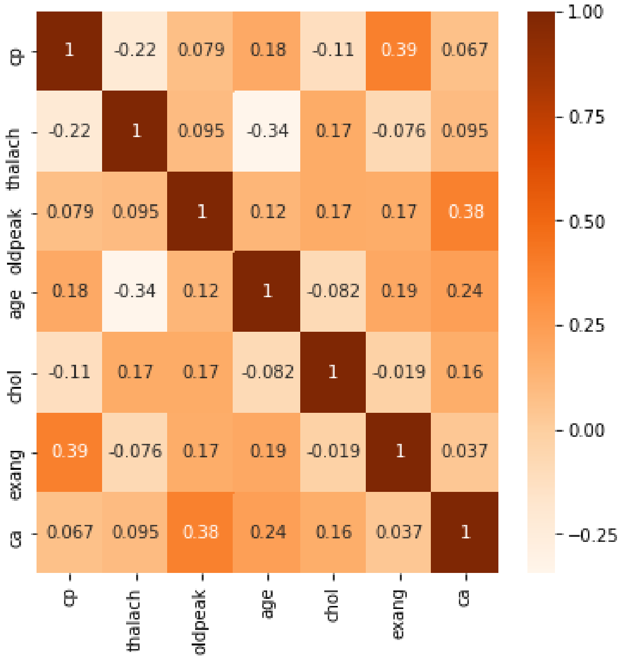
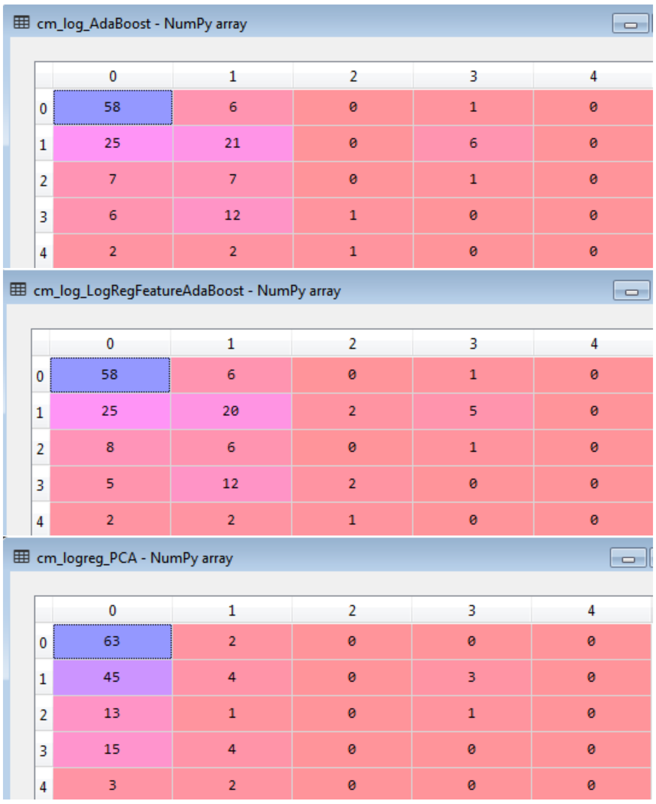


Fig 6. Features from Ada Boost Regressors

The dataset is applied to feature scaling and then fitted to kernel support vector machine classifier, logistic regression classifier, Naive bayes classifier, random forest classifier, decision tree classifier and KNN classifier and the obtained confusion matrix is shown in Fig. 7 – Fig. 12.



cm_log_AdaBoost - NumPy array

	0	1	2	3	4
0	58	6	0	1	0
1	25	21	0	6	0
2	7	7	0	1	0
3	6	12	1	0	0
4	2	2	1	0	0

cm_log_LogRegFeatureAdaBoost - NumPy array

	0	1	2	3	4
0	58	6	0	1	0
1	25	20	2	5	0
2	8	6	0	1	0
3	5	12	2	0	0
4	2	2	1	0	0

cm_logreg_PCA - NumPy array

	0	1	2	3	4
0	63	2	0	0	0
1	45	4	0	3	0
2	13	1	0	1	0
3	15	4	0	0	0
4	3	2	0	0	0

Fig. 7. Logistic Regression classifier

cm_log_KNN_AdaBoost - NumPy array

	0	1	2	3	4
0	49	14	2	0	0
1	33	14	3	2	0
2	8	5	1	1	0
3	11	6	1	1	0
4	5	0	0	0	0

cm_log_KNNFeatureAdaBoost - NumPy array

	0	1	2	3	4
0	52	12	1	0	0
1	20	23	7	2	0
2	6	7	1	1	0
3	5	6	4	4	0
4	2	1	1	1	0

cm_knn_PCA - NumPy array

	0	1	2	3	4
0	48	17	0	0	0
1	30	17	4	1	0
2	11	3	1	0	0
3	10	7	1	1	0
4	4	1	0	0	0

Fig.8. KNN classifier

cm_log_KSVM_AdaBoost - NumPy array

	0	1	2	3	4
0	64	1	0	0	0
1	48	1	3	0	0
2	13	1	1	0	0
3	16	2	1	0	0
4	5	0	0	0	0

cm_FeatureKSVM_AdaBoost - NumPy array

	0	1	2	3	4
0	57	8	0	0	0
1	22	27	0	3	0
2	8	7	0	0	0
3	5	10	2	2	0
4	1	2	2	0	0

cm_kernelsvm_PCA - NumPy array

	0	1	2	3	4
0	60	5	0	0	0
1	39	11	0	2	0
2	9	5	1	0	0
3	14	3	2	0	0
4	5	0	0	0	0

Fig. 9. Kernel SVM Classifier

cm_NaiveBayes_AdaBoost - NumPy array

	0	1	2	3	4
0	52	10	0	0	3
1	18	26	1	5	2
2	6	7	1	1	0
3	4	10	3	0	2
4	1	2	2	0	0

cm_FeatureNaiveBayes_AdaBoost - NumPy array

	0	1	2	3	4
0	52	10	0	0	3
1	18	26	1	5	2
2	6	7	1	1	0
3	4	10	3	0	2
4	1	2	2	0	0

cm_NB_PCA - NumPy array

	0	1	2	3	4
0	60	5	0	0	0
1	31	15	0	6	0
2	8	6	0	1	0
3	11	7	0	1	0
4	4	0	0	1	0

Fig. 10. Naïve Bayes Classifier

cm_RandomForest_AdaBoost - NumPy array

	0	1	2	3	4
0	49	14	0	2	0
1	21	19	7	5	0
2	5	7	2	1	0
3	6	5	5	3	0
4	2	1	0	2	0

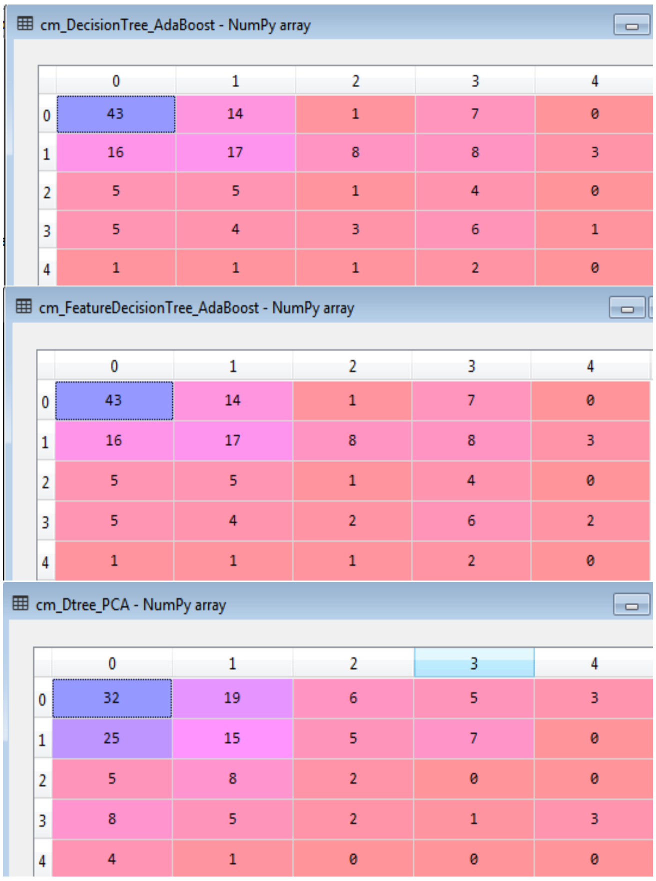
cm_FeatureRandomForest_AdaBoost - NumPy array

	0	1	2	3	4
0	50	13	0	2	0
1	21	19	7	5	0
2	5	7	2	1	0
3	7	4	5	2	1
4	2	1	0	2	0

cm_RandomForest_AdaBoost - NumPy array

	0	1	2	3	4
0	49	14	0	2	0
1	21	19	7	5	0
2	5	7	2	1	0
3	6	5	5	3	0
4	2	1	0	2	0

Fig. 11. Random Forest Classifier



cm_DecisionTree_AdaBoost - NumPy array

	0	1	2	3	4
0	43	14	1	7	0
1	16	17	8	8	3
2	5	5	1	4	0
3	5	4	3	6	1
4	1	1	1	2	0

cm_FeatureDecisionTree_AdaBoost - NumPy array

	0	1	2	3	4
0	43	14	1	7	0
1	16	17	8	8	3
2	5	5	1	4	0
3	5	4	2	6	2
4	1	1	1	2	0

cm_Dtree_PCA - NumPy array

	0	1	2	3	4
0	32	19	6	5	3
1	25	15	5	7	0
2	5	8	2	0	0
3	8	5	2	1	3
4	4	1	0	0	0

Fig. 12. Decision Tree Classifier

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 Ada 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	0.90	0.90	0.90

Table. 2 Recall Comparison for Ada 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	0.91	0.91	0.91

Table. 3 FScore Comparison for Ada 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	0.92	0.92	0.92

Table. 4 Accuracy Comparison for Ada 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	91	93	92

V. CONCLUSION

This paper makes an effort to predict the type of heart disease by considering the boosting of the parameters that are involved in prediction of the heart disease. The Ada Boost regressors were used to find the important features and attributes of the heart disease dataset. Experimental results show that, the Naïve bayes classifier is more effective with the precision of 0.90 for dataset with random boost, feature scaled and PCA. Experimental results show that, the Naïve bayes classifier is more effective with the recall of 0.91 for dataset with random boost, feature scaled and PCA. Experimental results show that, the Naïve bayes classifier is more effective with the Fscore of 0.92 for dataset with random boost, feature scaled and PCA. Experimental results show, the Naïve bayes classifier is more effective with the accuracy of 91% without random boost, 93% with random boosting and 92% with principal component analysis.

REFERENCES

1. Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.
2. Kannammal Sampathkumar," Analysis of Heart Disease Prediction using Various Machine Learning Techniques", Proceedings of the International Conference on Artificial Intelligence, Smart Grid and Smart City Applications, January 2019
3. Muthuvel, Marimuthu & Abinaya, M & Hariesh, K & Madhankumar, K & Pavithra, V, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach", International Journal of Computer Applications. 181. 975-8887. 10.5120/ijca2018917863.
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.