International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning

R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew

Abstract*: *In current marketing scenario, it is highly difficult to earn the high profit by satisfying the customers as well to increase to turn over of the company. For increasing the profit, the organizations are struggling to find a method to analyze their marketing strategy and to understand the customer's requirements. The main solution to increase the profit of any organization is to manufacture the limited and the needed goods based on the customer's needs and dislikes. For this, they need to find the customers behavior and the opinion regarding their products. This claims the usage of machine learning algorithms to predict and analyze the behavior of the customer. With this information scenario, we have extracted the wine data set from UCI Machine learning repository. The wine data set is analyzed to decide the dependent and independent variable. The dimensionality reduction is done by applying the ensembling methods. The feature importance of the various ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor and Gradient booster regressor. The extracted feature importance of the wine data set is fitted with logistic regression classifier to analyse the performance of the each ensembling methods. The metrics used for performance analysis are accuracy, precision, recall, and f-score. Experimental results shows that feature importance obtained from Ada Boost regressor fitted with logistic regression classifier is found to be effective with the accuracy of 94%, Precision of 0.95, Recall of 0.94 and FScore of 0.94 compared to other ensembling methods.

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

I. INTRODUCTION

The prediction of the customer behavior and their satisfaction are directly related to the growth of the profit and revenue of the organization. The company profit will generally be increased only when the customer buy the product constantly with full satisfaction. The single customer satisfaction may greatly influence the other customers in buying the product. So the organization must work hard to earn the customers satisfaction by developing the products based on their need and requirements. 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 survey is attempted to design by analyzing the arious wine data set attributes and the behavior patterns of the wine date set. This survey is used in predicting the customer's behavior so as to increase the profit of the organization. This is attempted with the motive of finding and analyzing the customer's intentions in buying the product. [1]

The ingredients in the wine product also greatly influence the people to buy the product and they attempted to find the essential ingredients by analyzing the feedback and the past history of product sales [2]. The composition of the mixing of the product also greatly influences the people in buying the product. They attempt to find the combination of the ingredients in forming the wine product and shown that the sales of the wine product got increased due to exact combination [3].

The model was created to develop each wine product and the standardization of the product is done based on the benchmark survey level of the wine consumers and the prediction is done with the benchmark data [4]. The factor analysis is done for the wine data set so as to reduce the number of features in the wine. This results in the data set with the reduced features so as to optimize the prediction with the minimum level of attributes and also they improvise the prediction accuracy. The simplified reduced principal component dataset were subjected to analyze the performance in increasing the profit of the company [5]. The product composition and the ingredients proposition in analyzing the wine quality and the profit was done based on the multi dimensional machine learning approaches which predicts the wine quality and the customer needs [6]. The use of feature selection, feature extraction and the classification that can be used to predict the wine quality is done [7]-[11].

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
R. Suguna*, Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
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.
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/)

III. PROPOSED WORK

In our proposed work, the wine data set is used for the segmentation of the customer group based on the need of their requirements. Our implementation in this paper is folded in four ways.

(i) Firstly, the connections between each of the attributes of the wine data set is depicted with the correlation pie representation.

(ii)Secondly, the important features are identified for each of the ensembling methods like Ada boost regressor, Random forest regressor, Extra Trees Regressor, Gradient booster regressor and Ada boost classifier.

(iii) Thirdly, the feature importance reduced data set from each of the ensembling methods are fitted to logistic regression classifier.

(iv) Fourth, the performance of the logistic regression classifier for each of the ensembling methods is analyzed by Precision, Recall, FScore and Accuracy.

A. System Architecture

The proposed system architecture of our work is shown in Fig. 1

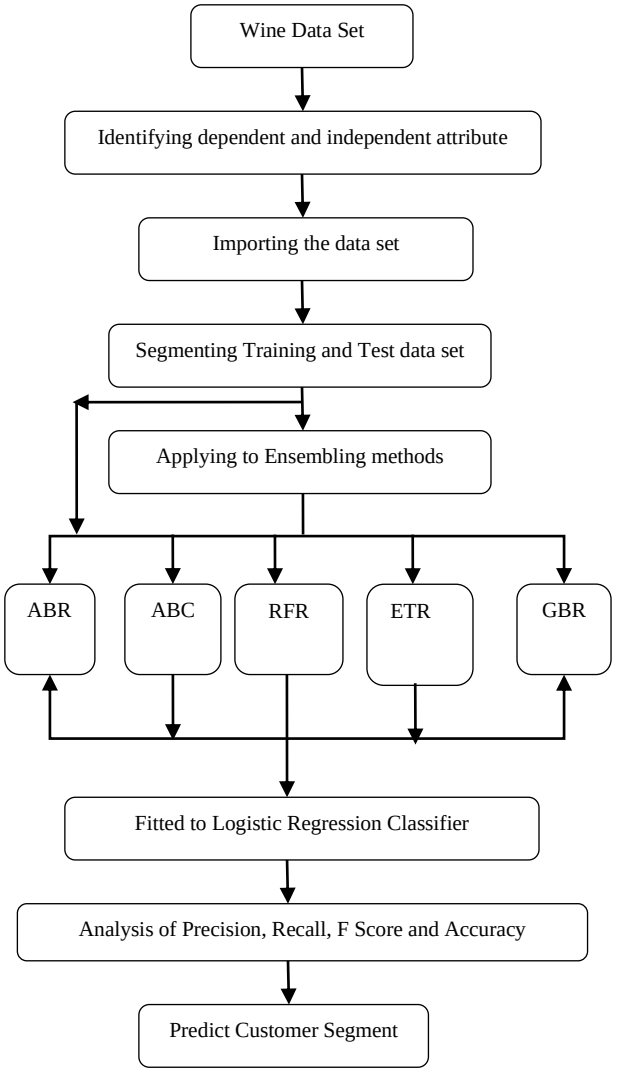


Fig. 1 System Architecture

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Segmentation of Customer Group

The Wine dataset is taken from UCL ML Repository for implementation with 13 independent attribute and 1 Customer Segment dependent attribute. The attribute are shown below.

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline Churn
14. Customer Segment - Dependent Attribute

The connections between each of the attributes of the wine data set are depicted with the correlation pie representation and are shown in Fig. 2.

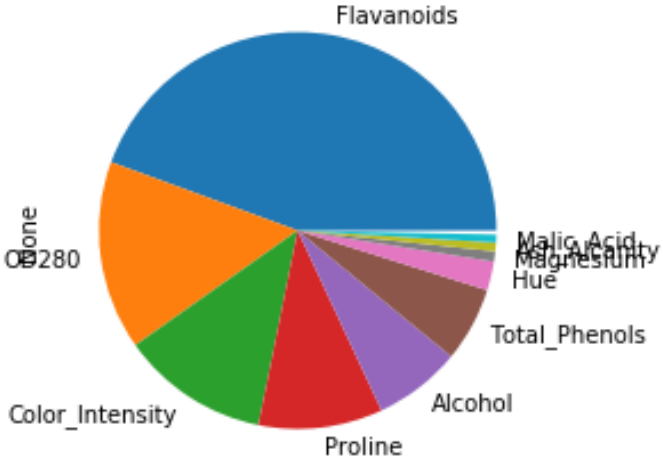
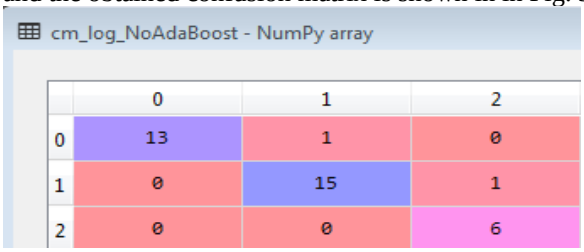


Fig. 2 Correlation Matrix Pie of Wine data set

The entire raw data set is fitted to logistic regression classifier and the obtained confusion matrix is shown in in Fig. 3.


cm_log_NoAdaBoost - NumPy array

	0	1	2
0	13	1	0
1	0	15	1
2	0	0	6

Fig. 3 Confusion Matrix for Logistic Regression classifier without applying Ensembling Methods

The correlation matrix of the important features obtained from the Ada Boost Regressor is shown in Fig. 4.

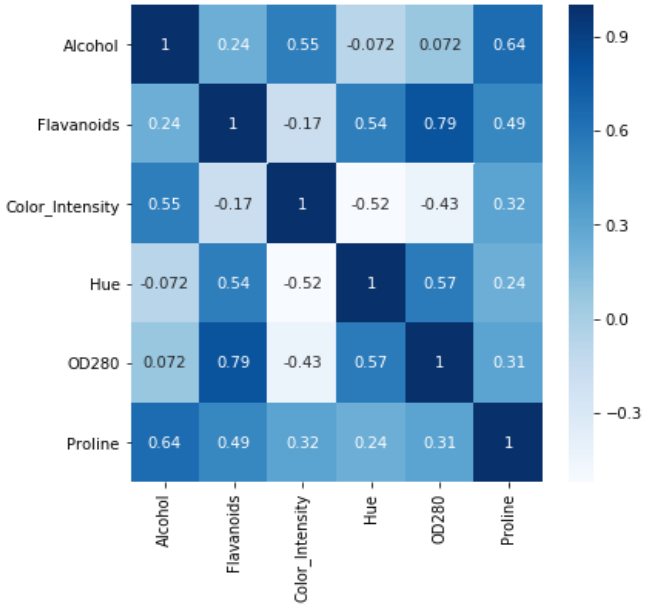


Fig. 4 Feature Importance Correlation Matrix of Ada Boost Regressor

The feature importance obtained from the Ada Boost Regressor is fitted to logistic regression classifier and the obtained confusion matrix is shown in in Fig. 5.

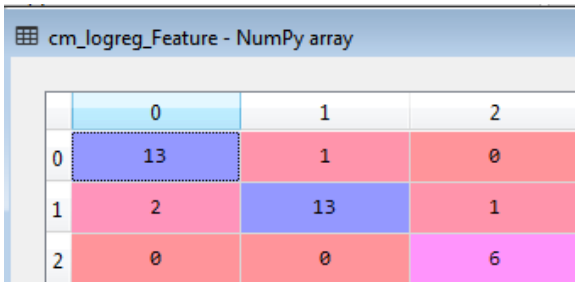


Fig. 5 Confusion Matrix for Logistic Regression with Ada Boost Regressor.

The correlation matrix of the important features obtained from the Random Forest Regressor is shown in Fig. 6.

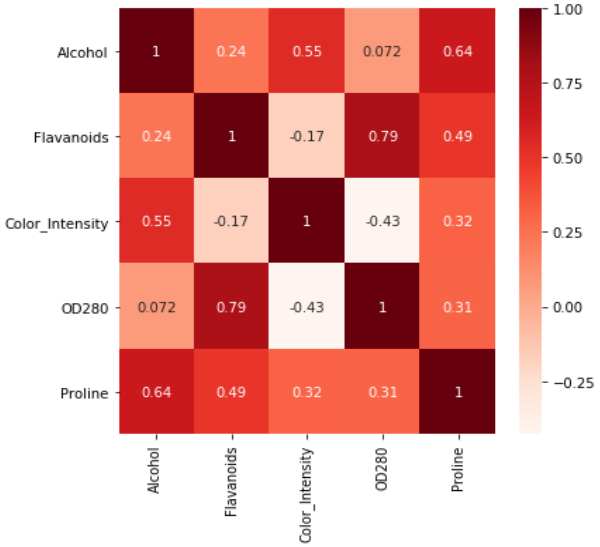


Fig. 6 Feature Importance Correlation Matrix of Random Forest Regressor

The feature importance obtained from the Random Forest Regressor is fitted to logistic regression classifier and the obtained confusion matrix is shown in in Fig. 7.

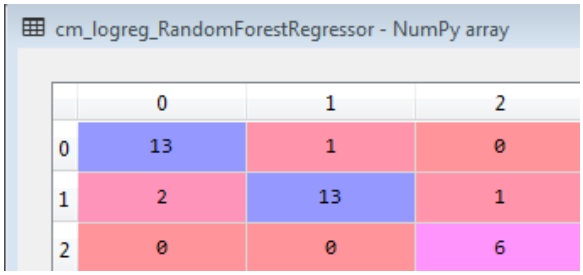


Fig. 7 Confusion Matrix for Logistic Regression with Random Forest Regressor

The correlation matrix of the important features obtained from the Extra Trees Regressor is shown in Fig. 8.

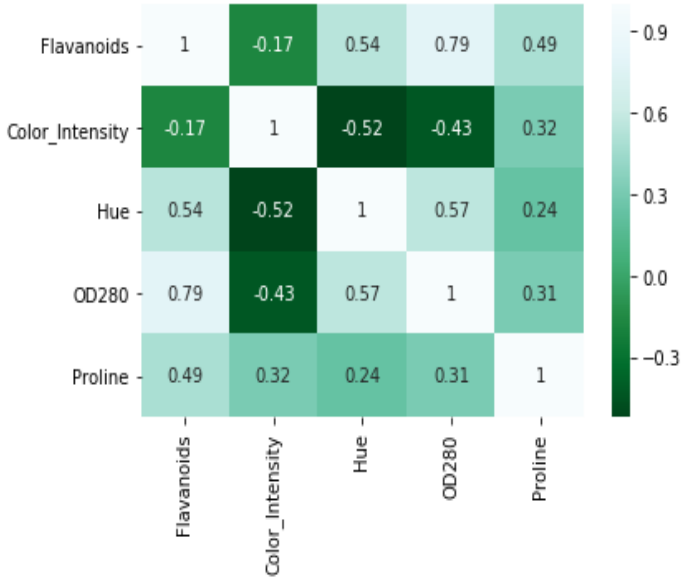


Fig. 8 Feature Importance Correlation Matrix of Extra Trees Regressor

The feature importance obtained from the Extra Trees Regressor is fitted to logistic regression classifier and the obtained confusion matrix is shown in in Fig. 9.

Fig. 9 Confusion Matrix for Logistic Regression with Extra Trees Regressor

The feature importance obtained from the Gradient Boosting Regressor is fitted to logistic regression classifier and the obtained confusion matrix is shown in in Fig. 10.

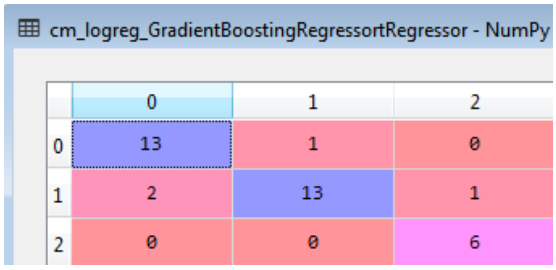


Fig. 10 Confusion Matrix for Logistic Regression with Gradient Boosting Regressor

The correlation matrix of the important features obtained from the Ada Boost Classifier is shown in Fig. 11.

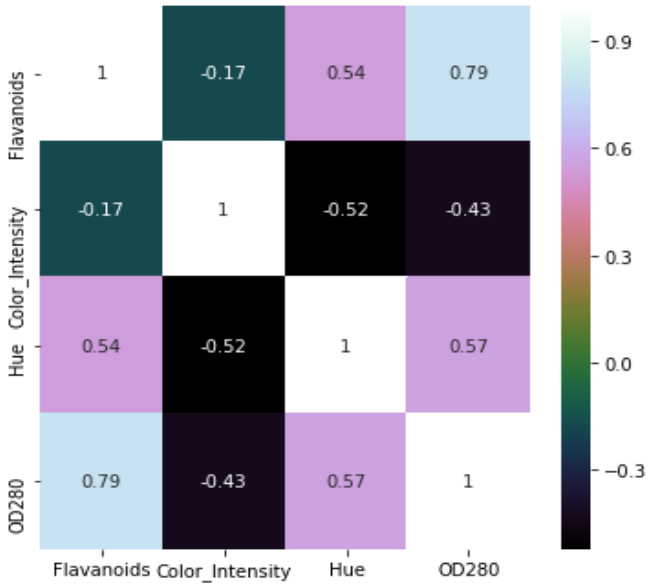


Fig. 11 Feature Importance Correlation Matrix of Ada Boost Classifier

The feature importance obtained from the Ada Boost Classifier is fitted to logistic regression classifier and the obtained confusion matrix is shown in in Fig. 12.

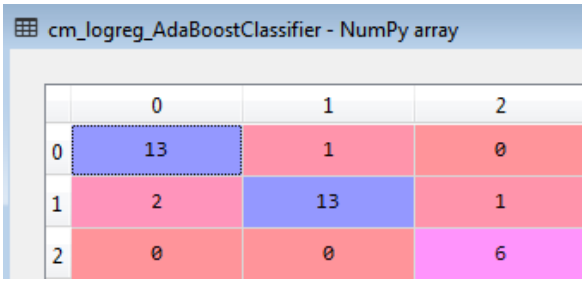


Fig. 12 Confusion Matrix for Logistic Regression with Ada Boost Classifier.

The feature importance reduced data set from each of the ensembling methods are fitted to logistic regression classifier.

The performance of the logistic regression classifier for each of the ensembling methods is analyzed by Precision, Recall, FScore and Accuracy and is shown in the Table. 1 and the Table. 2.

Table. 1 Performance Comparison of Precision, Recall and FScore for Logistic Regression Classifier

Ensembling Methods	Fitted to Logistic Regression Classifier		
	Precision	Recall	FScore
Ada Boost Regressor	0.95	0.94	0.94
Random Forest Regressor	0.89	0.89	0.89
Extra Trees Regressor	0.88	0.88	0.89
Gradient Boost Regressor	0.87	0.88	0.87
Ada Boost Classifier	0.88	0.88	0.89

Table. 2 Performance Comparison of Accuracy for Logistic Regression Classifier

Ensembling Methods	Accuracy for Logistic Regression Classifier (%)
Ada Boost Regressor	94
Random Forest Regressor	89
Extra Trees Regressor	88
Gradient Boost Regressor	87
Ada Boost Classifier	89

The Performance analysis of the metrics like precision, recall, FScore and Accuracy is shown in Fig. 13-Fig. 16.

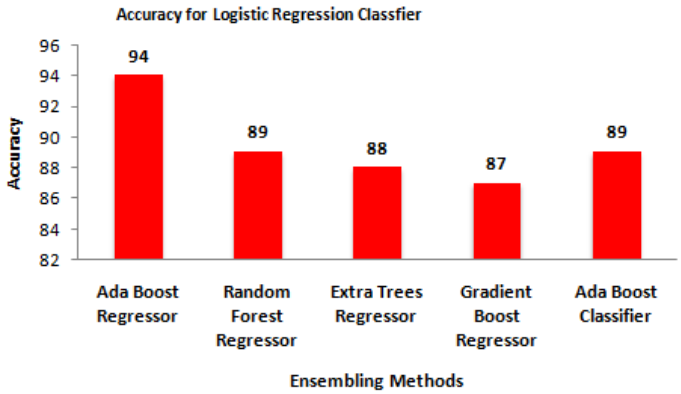


Fig. 13. Accuracy Analysis for Ensembling Methods

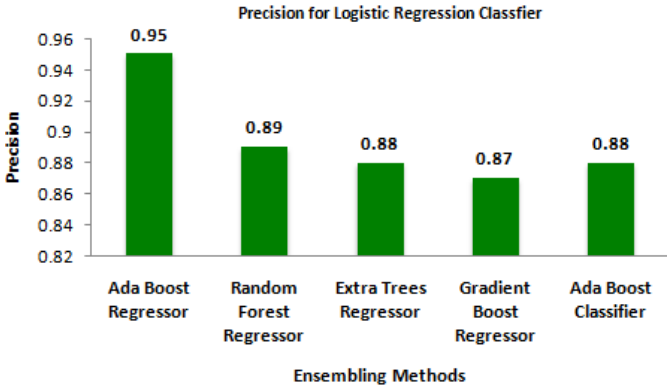


Fig. 14. Precision Analysis for Ensembling Methods

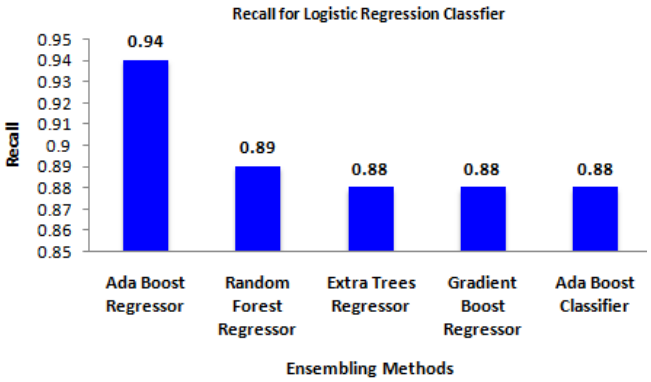


Fig. 15. Recall Analysis for Ensembling Methods

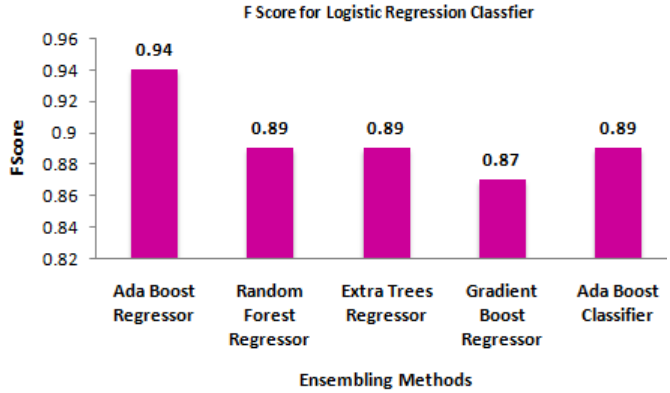


Fig. 16. FScore Analysis for Ensembling Methods

V. CONCLUSION

This paper attempts to analyze the segmentation of the customer group based on the attributes distribution in the wine data set. The dimensionality reduction is done by applying the ensembling methods. The feature importance of the various ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor and Gradient booster regressor. The extracted feature importance of the wine data set is fitted with logistic regression classifier to analyse the performance of the each ensembling methods. The metrics used for performance analysis are accuracy, precision, recall, and f-score. Experimental results shows that feature importance obtained from Ada Boost regressor fitted with logistic regression classifier is found to be effective with the accuracy of 94%, Precision of 0.95, Recall of 0.94 and FScore of 0.94 compared to other ensembling methods.

REFERENCES

1. D. Veena Parboteeah, D. Christopher Taylor, and A. Nelson Barber, “Exploring impulse purchasing of wine in the online environment”, Journal of Wine Research., vol. 27, no. 4, 2016, pp. 322-339.
2. Hyojin Kim, and A. Mark Bonn, “The Moderating Effects of Overall and Organic Wine Knowledge on Consumer Behavioral Intention”, Scandinavian Journal of Hospitality and Tourism., vol. 15, no. 3, 2015 pp. 295-310.
3. Johan Bruwer, Nicole Burrows, Sylvia Chaumont, Elton Li, and Anthony Saliba, “Consumer involvement and associated behaviour in the UK high-end retail off-trade wine market”, The International Review of Retail, Distribution and Consumer Research., vol. 24, no. 2, 2014, pp. 145-165.
4. Johan Bruwer, Justin Cohen, and Kathleen Kelley, "Wine involvement interaction with dining group dynamics, group composition and consumption behavioural aspects in USA restaurants", International Journal of Wine Business Research., vol. 3, no.1, 2019, pp.12-28.
5. Johan Bruwer, Polymeros Chrysochou, and Isabelle Lesschaeve., “Consumer involvement and knowledge influence on wine choice cue utilization”. British Food Journal., vol. 119, no. 4, 2017,pp. 830-844.
6. Y. Subba Reddy and Prof. P. Govindarajulu, “An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set”, International Journal of Computer Science and Network Security., vol..17, no. 10, October 2017.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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. DOI:10.1080/09720510.2019.1609729
12. 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.
13. 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.