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

Regressor Fitting Of Feature Importance For Customer Segment Prediction With Ensembling Schemes Using Machine Learning

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

Abstract*: Prediction of client behavior and their feedback remains as a challenging task in today's world for all the manufacturing companies. The companies are struggling to increase their profit and annual turnover due to the lack of exact prediction of customer like and dislike. This leads to the accomplishment of machine learning algorithms for the prediction of customer demands. This paper attempts to identify the important features of the wine data set extracted from UCI Machine learning repository for the prediction of customer segment. The important features are extracted for the various ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor, Gradient booster regressor. The extracted feature importance of each of the ensembling methods is then fitted with logistic regression to analyze the performance. The same extracted feature importance of each of the ensembling methods are subjected to feature scaling and then fitted with logistic regression to analyze the performance. The Performance analysis is done with the performance metric such as Mean Squared error (MSE), Mean Absolute error (MAE), R2 Score, Explained Variance Score (EVS) and Mean Squared Log Error (MSLE). Experimental results shows that after applying feature scaling, the feature importance extracted from the Extra Tree Regressor is found to be effective with the MSE of 0.04, MAE of 0.03, R2 Score of 94%, EVS of 0.9 and MSLE of 0.01 as compared to other ensembling methods.

***Index Terms*: Machine Learning, Mean Squared error, Mean Absolute error, R2 Score, Explained Variance Score and Mean Squared Log Error.**

I. INTRODUCTION

Generally the dataset in the market have a lot of attributes. The single dependent variable of the dataset is predicted by the occurrence of one or more independent variables. However, the dependent variable does not need the existence of all independent variable for its predicted. Some of the independent variable are not at all involved in the prediction of the target variable. So it is very essential to find the important features of the machine learning dataset so as to predict the value of the dependent variable with high accuracy. 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 chemical samples and its proposition is needed to predict the quality of wine. Due to the change in the mixing of the chemicals and their existence in the wine, the quality of wine greatly changes. Based on the quality of the wine, the customers prefer the product. The machine learning models can be built to find the exact combination of the chemicals to be added based on the customers behavior. The machine learning models like Linear Regression, Decision Trees and Artificial Neural Networks are used to predict the customer behavior that helps in finding the needed features to understand the customer's behavior and demand [1].

The data mining techniques are used to predict the customers need and their behavior in choosing the wine. The statistics that are involved in the data mining techniques can find the exact combination of the independent variables that are present in the dataset [2].

The customer relationship management is greatly needed for any business to survive in the current market world. The utilization charge of wine was evaluated using various factors like such as manufactured goods involvement, biased awareness, delicate qualities and socio demography [3].

Due to the growth in the online shopping, the customers wish to buy the high quality wine through online web portal shopping. In this scenario, the customers just view the quality of the wine only through the ingredients present in the wine [4]. The various wine brands has worth in their improvement and the current market is highly competitive [5].

A critical review on various feature selection, feature extraction methods, classification methods and the performances parameters are examined for predicting the wine quality [6]-[10].

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.
Rincy Merlin Mathew, Lecturer, Department of Computer Science, College of Science and Arts, Khamis Mushayt, King Khalid university, Abha, Asir, Saudi Arabia.
R. Suguna, 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 wine data set is applied to extract the important features using various ensembling methods.. Our implementation in this paper is folded in five ways.

(i) Firstly, the analysis of correlation matrix for the entire features of the wine data set

(ii)Secondly, extracting the feature importance of the ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor, Gradient booster regressor.

(iii) Thirdly, extracted important features of the various ensembling methods are fitted to logistic regression methods.

(iv) Fourth, the extracted important features of the various ensembling methods are subjected to feature scaling and then fitted to logistic regression methods.

(v) Fifth, the performance analysis of the feature importance of various ensembling methods are done by Mean Squared error, Mean Absolute error, R2 Score, Explained Variance Score and Mean Squared Log Error.

A. System Architecture

The propose architecture of our work is shown in Fig. 1

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Customer Segment Prediction

The Wine dataset is extracted from UCL Machine Learning Repository is used for implementation with 13 independent attribute and 1 Customer Segment dependent attribute. The correlation matrix of the wine data set is depicted in Fig .2. and is used to identify the relationship between the features.

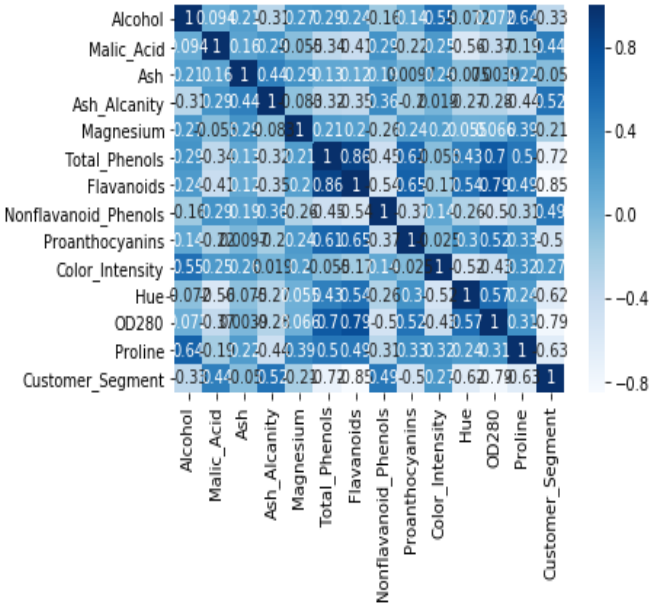


Fig. 2 Correlation Matrix of Wine data set

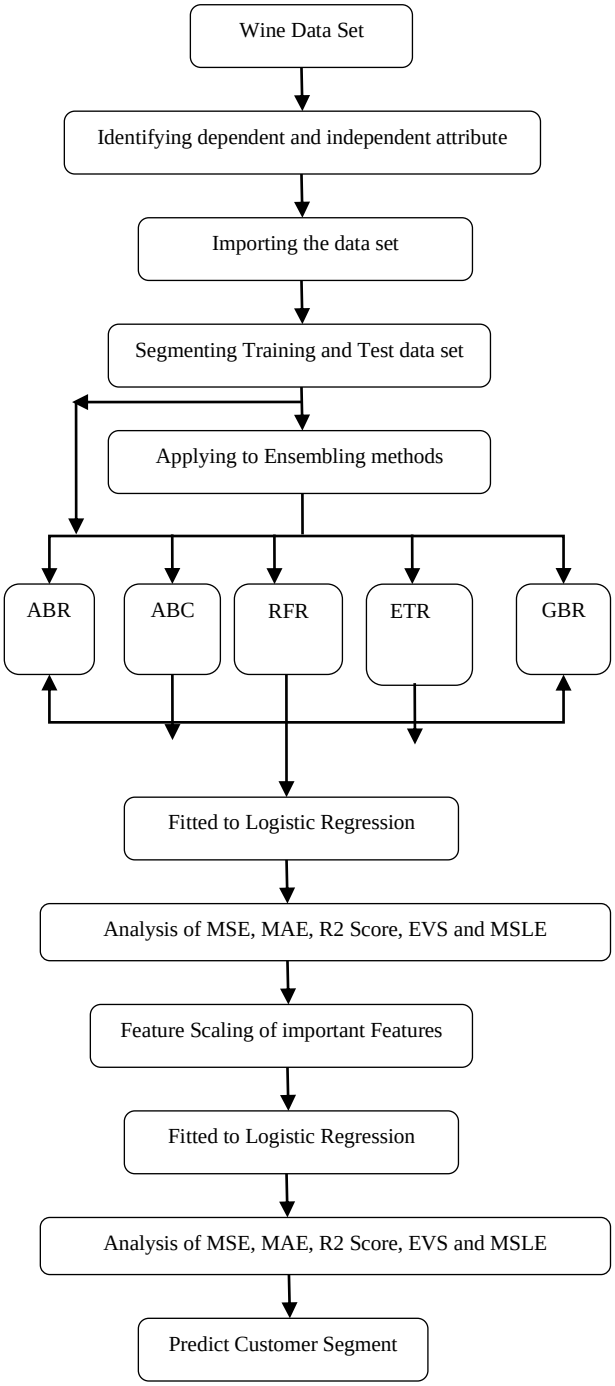


Fig. 1 System Architecture

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 feature importance variables of the wine data set extracted from the Ada boost regressor is shown in Fig. 3.

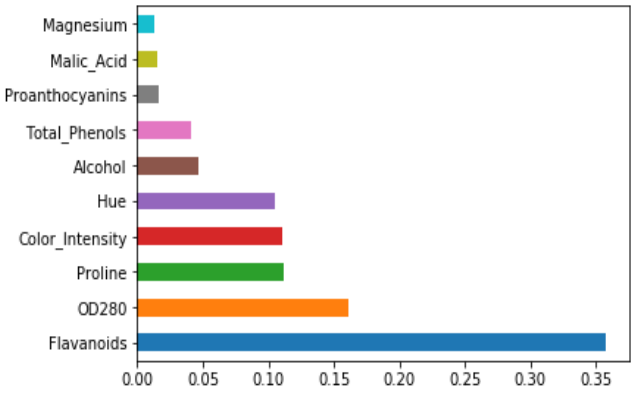


Fig. 3 Feature Importance of Ada Boost Regressor

The distribution of high feature component along with their variance values for the Ada boost regressor and Ada boost classifier is shown in Fig. 4.

AdaBoost_featureimportances - Serie

Index	0
Alcohol	0.0467108
Malic_Acid	0.0155959
Ash	0.00621346
Ash_Alcanity	0.010704
Magnesium	0.0127728
Total_Phenols	0.0410398
Flavanoids	0.357936
Nonflavanoid…	0.00372344
Proanthocyan…	0.0165147
Color_Intens…	0.110811
Hue	0.105176
OD280	0.16068
Proline	0.112122

AdaBoostClassifier_feature_importan

Index	0
Alcohol	0
Malic_Acid	0
Ash	0
Ash_Alcanity	0
Magnesium	0
Total_Phenols	0
Flavanoids	0.5
Nonflavanoid…	0
Proanthocyan…	0
Color_Intens…	0.02
Hue	0.46
OD280	0.02
Proline	0

Fig. 4 Important Features of Ada Boost Regressor and Classifier

The feature importance variables of the wine data set extracted from the Ada boost classifier is shown in Fig. 5.

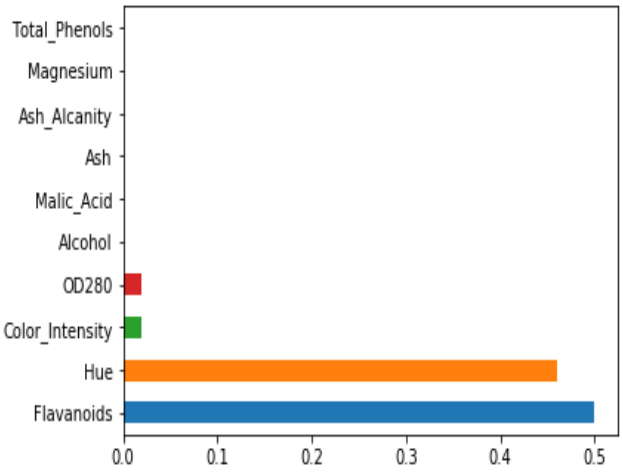


Fig. 5 Feature Importance of Ada Boost Classifier

The distribution of high feature component along with their variance values for the Random forest regressor and Extra tress regressor is shown in Fig. 6.

RandomForestRegressor_feature_imp

Index	0
Alcohol	0.0755097
Malic_Acid	0.00210284
Ash	0.00056346
Ash_Alcanity	0.00309196
Magnesium	0.0038305
Total_Phenols	0.000648047
Flavanoids	0.357601
Nonflavanoid…	0.000400574
Proanthocyan…	0.00299872
Color_Intens…	0.110102
Hue	0.00887572
OD280	0.314037
Proline	0.120238

ExtraTreesRegressor_feature_importa

Index	0
Alcohol	0.0594758
Malic_Acid	0.00910158
Ash	0.00633867
Ash_Alcanity	0.014938
Magnesium	0.0111067
Total_Phenols	0.0532582
Flavanoids	0.315258
Nonflavanoid…	0.00359407
Proanthocyan…	0.0113401
Color_Intens…	0.0598497
Hue	0.0696181
OD280	0.260374
Proline	0.125747

Fig. 6 Important Features of Random Forest and Extra Trees Regressor

The feature importance variables of the wine data set extracted from the Random forest regressor is shown in Fig.7.

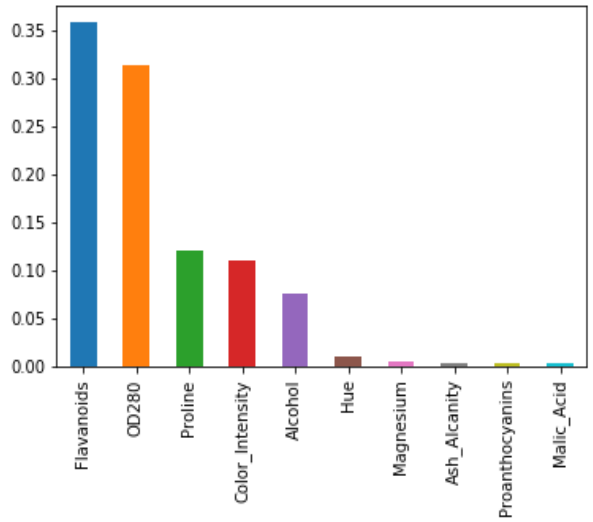


Fig. 7 Feature Importance of Random Forest Regressor

The feature importance variables of the wine data set extracted from the Extra Trees regressor is shown in Fig.8.

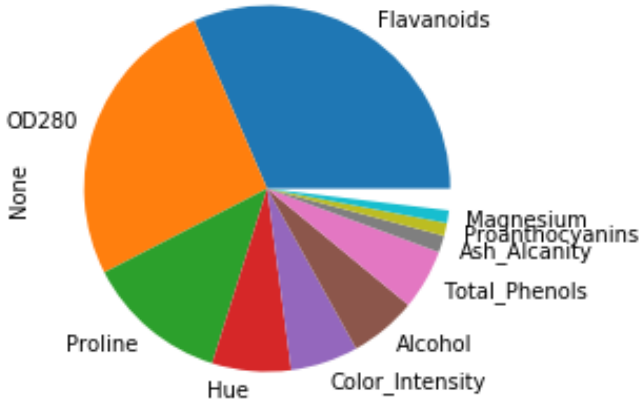


Fig. 8 Feature Importance of Extra Trees Regressor

The feature importance variables of the wine data set extracted from the Gradient Boosting regressor is shown in Fig. 9.

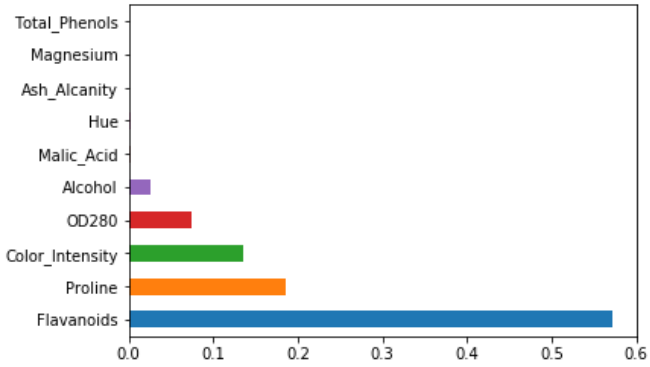
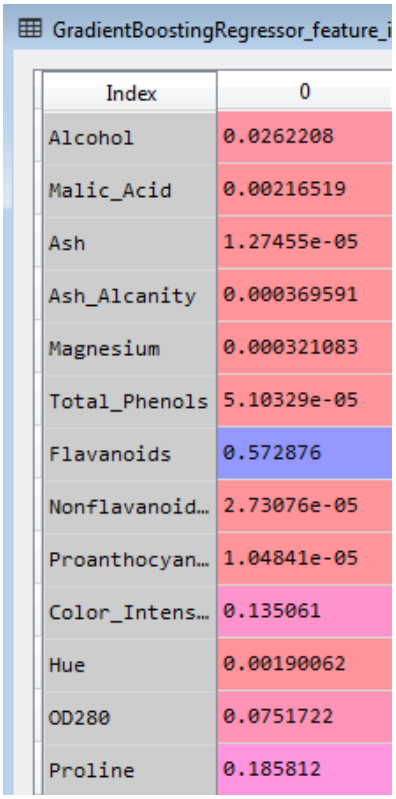


Fig. 9 Feature Importance of Gradient Boosting Regressor

The distribution of high feature component along with their variance values for the Gradient Boosting regressor is shown in Fig. 10.


GradientBoostingRegressor_feature_i

Index	0
Alcohol	0.0262208
Malic_Acid	0.00216519
Ash	1.27455e-05
Ash_Alcanity	0.000369591
Magnesium	0.000321083
Total_Phenols	5.10329e-05
Flavanoids	0.572876
Nonflavanoid…	2.73076e-05
Proanthocyan…	1.04841e-05
Color_Intens…	0.135061
Hue	0.00190062
OD280	0.0751722
Proline	0.185812

Fig. 10 Important Features of Gradient Boosting Regressor

The extracted important features of the various ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor, Gradient booster regressor are then fitted to logistic regression methods and the performance is analyzed. The performance analysis of the feature importance of various ensembling methods are done by Mean Squared error, Mean Absolute error, R2 Score, Explained Variance Score and Mean Squared Log Error. The performance metric comparison is shown in the Table. 1 and Table. 2.

Table. 1 Performance Comparison of MSE, MAE and R2 Score for various Ensembling before Feature Scaling

Ensembling Methods	Fitting to Logistic Regression Before Feature Scaling		
	MSE	MAE	R2 Score
AdaBoost Regressor	0.11	0.11	0.78
Ada boost Classifier	0.19	0.19	0.61
Random Forest Regressor	0.12	0.12	0.79
Extra Trees Regressor	0.08	0.08	0.83
Gradient Boosting Regressor	0.11	0.11	0.78

Table. 2 Performance Comparison of EVS and MSLE Score for various Ensembling before Feature Scaling

Ensembling Methods	Fitting to Logistic Regression Before Feature Scaling	
	EVS	MSLE
AdaBoost Regressor	0.78	0.15
Ada boost Classifier	0.69	0.02
Random Forest Regressor	0.79	0.16
Extra Trees Regressor	0.83	0.01
Gradient Boosting Regressor	0.78	0.15

The extracted important features of the various ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor, Gradient booster regressor are subjected to feature scaling and then fitted to logistic regression methods. The performance analysis of the feature importance of various ensembling methods are done by Mean Squared error, Mean Absolute error, R2 Score, Explained Variance Score and Mean Squared Log Error. The performance metric comparison is shown in the Table. 3 and Table. 4.

Table. 3 Performance Comparison of MSE, MAE and R2 Score for various Ensembling after Feature Scaling

Ensembling Methods	Fitting to Logistic Regression after Feature Scaling		
	MSE	MAE	R2 Score
AdaBoost Regressor	0.27	0.27	0.94
Ada boost Classifier	0.16	0.16	0.69
Random Forest Regressor	0.26	0.26	0.93
Extra Trees Regressor	0.04	0.03	0.94
Gradient Boosting	0.27	0.27	0.93

Table. 4 Performance Comparison of EVS and MSLE Score for various Ensembling after Feature Scaling

Ensembling Methods	Fitting to Logistic Regression after Feature Scaling	
	EVS	MSLE
AdaBoost Regressor	0.74	0.02
Ada boost Classifier	0.69	0.03
Random Forest Regressor	0.73	0.02
Extra Trees Regressor	0.93	0.01
Gradient Boosting Regressor	0.84	0.02

V. CONCLUSION

This paper attempts to predict the customer behaviour by extracting the important features from the wine data set. The correlation matrix of the wine data set is recognized between each attributes in the wine data set. The extracted feature importance of each of the ensembling methods like Ada boost regressor, Ada boost classifier, Random forest regressor, Extra Trees Regressor, Gradient booster regressor is then fitted with logistic regression to analyze the performance. The same extracted feature importance of each of the ensembling methods are subjected to feature scaling and then fitted with logistic regression to analyze the performance. The Performance analysis is done with the performance metric such as MSE, MAE, R2 Score, EVS and MSLE. Experimental results shows that after applying feature scaling, the feature importance extracted from the Extra Tree Regressor is found to be effective with the MSE of 0.04, MAE of 0.03, R2 Score of 94%, EVS of 0.9 and MSLE of 0.01 as compared to other ensembling methods.

REFERENCES

1. Jorge Ribeiro , Jose Neves , Juan Sanchez , Manuel Delgado , Jose Machado, and Paulo Novais, "1. Wine Vinification prediction using Data Mining tools", Computing and Computational Intelligence, 2016
2. T. Afolabi Ibukun, Olufunke Oladipupo, E. Rowland Worlu, and O. Akinyemi," An Open Access Journal Available Online A Systematic Review of Consumer Behaviour Prediction Studies", 2. Covenant Journal of Business & Social Sciences., vol. 7, no. 1, June 2016.
3. David Cox, "Predicting Consumption, Wine Involvement and Perceived Quality of Australian Red Wine", Journal of Wine Research., vol. 20, no. 3, 2009, pp. 209-229.
4. Constanza Bianchi, "Consumer Brand Loyalty in the Chilean Wine Industry", Journal of Food Products Marketing., vol. 21, no. 4, 2015, pp. 442-460.
5. Austin Waters, and Risto Miikkulainen, "GRADE: Machine Learning Support for Graduate Admissions", Association for the Advancement of Artificial Intelligence, Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference., 2013
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
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. 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.