International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Count Vectorized Spam and Ham Discernment of Short Message Service using Machine Learning Classification

M. Shyamala Devi, Kamma Rahul, Manubolu Satheesh, Koruprolu Rajasekhar, Pittala Ganesh Kumar

Abstract*: With the growing volume and the amount of spam message, the demand for identifying the effective method for spam detection is in claim. The growth of mobile phone and Smartphone has led to the drastic increase in the SMS spam messages. The advancement and the clean process of mobile message servicing channel have attracted the hackers to perform their hacking through SMS messages. This leads to the fraud usage of other accounts and transaction that result in the loss of service and profit to the owners. With this background, this paper focuses on predicting the Spam SMS messages. The SMS Spam Message Detection dataset from KAGGLE machine learning Repository is used for prediction analysis. The analysis of Spam message detection is achieved in four ways. Firstly, the distribution of the target variable Spam Type the dataset is identified and represented by the graphical notations. Secondly, the top word features for the Spam and Ham messages in the SMS messages is extracted using Count Vectorizer and it is displayed using spam and Ham word cloud. Thirdly, the extracted Counter vectorized feature importance SMS Spam Message detection dataset is fitted to various classifiers like KNN classifier, Random Forest classifier, Linear SVM classifier, Ada Boost classifier, Kernel SVM classifier, Logistic Regression classifier, Gaussian Naive Bayes classifier, Decision Tree classifier, Extra Tree classifier, Gradient Boosting classifier and Multinomial Naive Bayes classifier. Performance analysis is done by analyzing the performance metrics like Accuracy, FScore, Precision and Recall. The implementation is done by python in Anaconda Spyder Navigator. Experimental Results shows that the Multinomial Naive Bayes classifier have achieved the effective prediction with the precision of 0.98, recall of 0.98, FScore of 0.98 , and Accuracy of 98.20%..

***Index Terms*: Machine Learning, Recall, FScore, Accuracy and AUC Score.**

I. PREAMBLE

In machine learning, the prediction of spam short message service message detection is done either by regression or classification process. The entire nation is moving towards usage of mobile and smart phone due to the technological growth. Due to mobile usage, it is easy to hack the mobile numbers of the people exposing the subscription at high risk.

The paper is planned in order to explain the existing details with Section 2 followed by the proposed work in the Section 3. Implementation and the performance analysis is discussed in Section 4 followed by the conclusion of the paper in Section 5.

II. RELATED WORK

A. Literature Survey

The dimensionality reduction can be done by the feature extraction and selection and is considered in predicting the target variable [1]. The general principles, basic idea and the benchmark level is used for analyzing the target variable [2]. The prediction of the target variable for SMS spam messages is done with the classification methods and it is used to categorize the class of transaction using Rule based approach [3]. The clustering and classification analysis is used for predicting the SMS spam messsage transaction [4]. The analysis of the whole sms spam data is needed for predicting the fraud detection and the machine learning approaches can be used to implement this [5]. Several data mining tools and approaches can be used for predicting the credit card fraud detection. The manual computation of detecting the fraud credit card online transaction detection is a tedious and time consuming process and it lead to impractical condition [6]. The fraud in the credit card transaction can be due to inner and outer environment and the fraud may be due to the credit card stole and unusual way of handling the online transaction [7]. The machine learning feature selection and feature extraction methods can be used for the prediction of any factor in different application can be learnt through this article [8] – [21].

Manuscript published on November 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.
Kamma Rahul, 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.
Manubolu Satheesh, 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.
Koruprolu Rajasekhar, 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.
Pittala Ganesh Kumar, 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.

III. PROPOSED WORK

In this paper, we have used machine learning classification algorithm for predicting the SMS Spam message detection. Our contribution of predicting SMS Spam message transaction is done in four ways.

(i) Firstly, the distribution of the target variable Spam Type the dataset is identified and represented by the graphical notations.

(ii) Secondly, the top word features for the Spam and Ham messages in the SMS messages is extracted using Count Vectorizer and it is displayed using word cloud.

(iii) Thirdly, the extracted Counter vectorized feature importance SMS Spam Message detection dataset is fitted to various classifiers like KNN classifier, Random Forest classifier, Linear SVM classifier, Ada Boost classifier, Kernel SVM classifier, Logistic Regression classifier, Gaussian Naive Bayes classifier, Decision Tree classifier, Extra Tree classifier, Gradient Boosting classifier and Multinomial Naive Bayes classifier

(iv) Performance analysis is done by analyzing the performance metrics like Accuracy, FScore, Precision and Recall.

A. System Architecture

The overall design of this paper is shown in Fig. 1

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Data Set Information

The SMS Spam Message detection from KAGGLE Machine Learning database warehouse is used for execution with 1 independent attribute and 1 Spam Type Class dependent attribute with 5572 number of rows and they are as follows,

(1) Sentence

(2) Spam Type (Target- Dependent Attribute)

SMS Spam Message detection Data Set is implemented to analyze the target distribution of fraud class and is shown in Fig. 2.

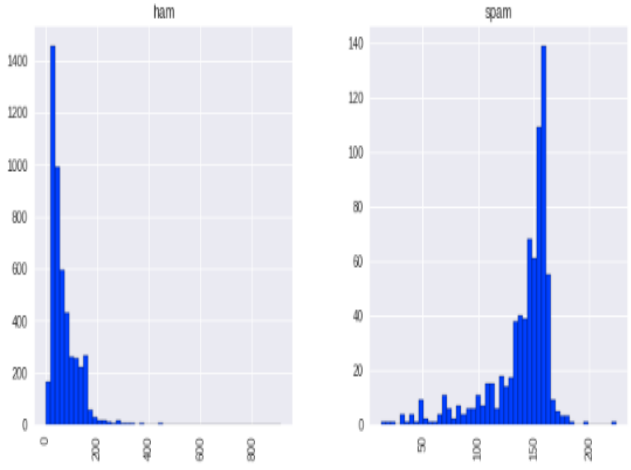

Fig. 2 Dataset Target division

B. Prediction of SMS Spam Message detection

SMS Spam Message detection Data Set is implemented to analyze the amount of target Spam and Ham distribution of and is shown in Fig. 3.

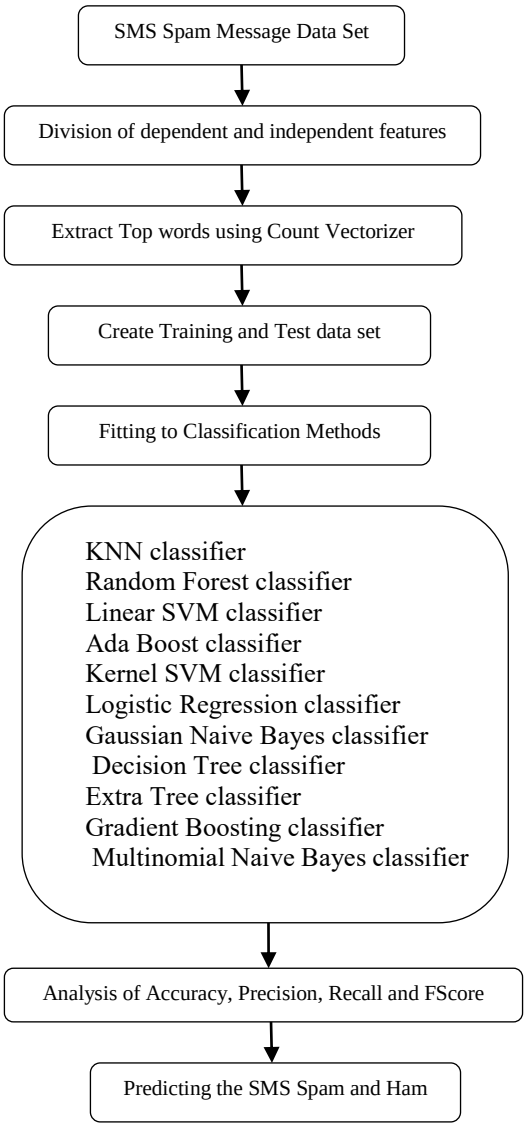

Fig. 1 System Architecture

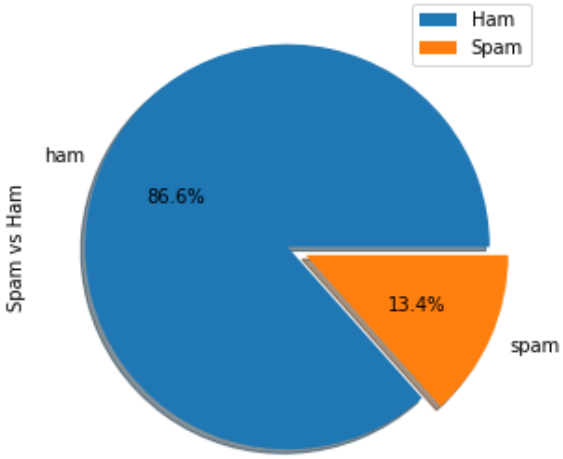

Fig. 3 Spam and Ham distribution of Dataset

Retrieval Number: D7287118419/2019©BEIESP
DOI:10.35940/ijrte.D7287.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The SMS Spam Message detection Data Set is subjected to find the top word features for the Spam and Ham messages in the SMS messages and is extracted using Count Vectorizer and it is displayed using word cloud and is shown in Fig. 4 – Fig. 5.

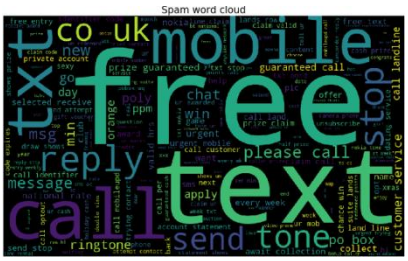

Fig. 4. Spam Word Cloud

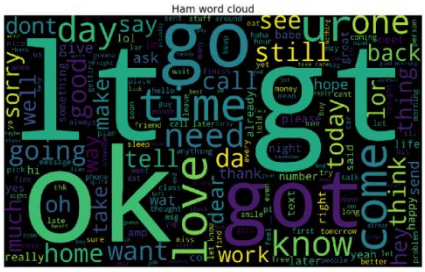

Fig. 5. Ham Word Cloud

The length of the Spam and Ham messages of SMS Spam Message detection Data Set is shown in Fig. 6 - Fig 7.

```
Average length of spam messages:
  138.8661311914324 characters
Average length of ham messages:
 71.02362694300518 characters
```

Fig. 6 Spam and Ham message length in the Data Set

The extracted Counter vectorized feature importance SMS Spam Message detection dataset is fitted to various classifiers like KNN classifier, Random Forest classifier, Linear SVM classifier, Ada Boost classifier, Kernel SVM classifier, Logistic Regression classifier, Gaussian Naive Bayes classifier, Decision Tree classifier, Extra Tree classifier, Gradient Boosting classifier and Multinomial Naive Bayes classifieris shown in Fig 8- Fig. 18.

```
In [25]: df1['Words'].valu
Out[25]:
           index  Words
0              u   1212
1           call    606
2            get    397
3             ur    385
4             gt    318
5             lt    316
6             ok    292
7           free    288
8             go    286
9           know    261
10          like    245
11          good    245
12           day    242
13           got    240
14          come    230
15          time    220
16          love    209
17          send    199
```

Fig. 7 Spam and Ham message length

cm_logreg_Count - NumPy array

	0	1
0	962	3
1	24	126

Fig. 8. Logistic Regression Confusion Matrix

cm_knn_Count - NumPy array

	0	1
0	958	7
1	79	71

Fig. 9. KNN Confusion Matrix

cm_Linearsvm_Count - NumPy array

	0	1
0	961	4
1	23	127

Fig. 10. Linear SVM Confusion Matrix

cm_kernelsvm_Count - NumPy array

	0	1
0	911	54
1	65	85

Fig. 11. Kernel SVM Confusion Matrix

cm_NB_GaussianNB_Count - NumPy a

	0	1
0	790	175
1	15	135

Fig. 12. Gaussian Naive Bayes Confusion Matrix

cm_Dtree_count - NumPy array

	0	1
0	952	13
1	21	129

Fig. 13. Decision Tree Confusion Matrix

Retrieval Number: D7287118419/2019©BEIESP
DOI:10.35940/ijrte.D7287.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

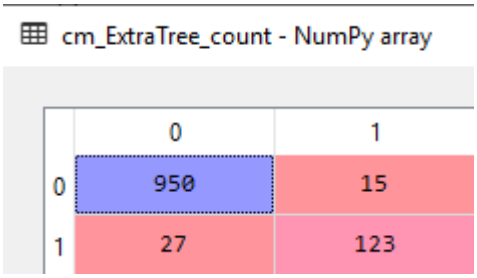

Fig. 14. Extra Tree Confusion Matrix

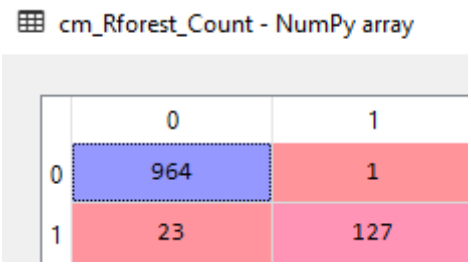

Fig. 15. Random Forest Confusion Matrix

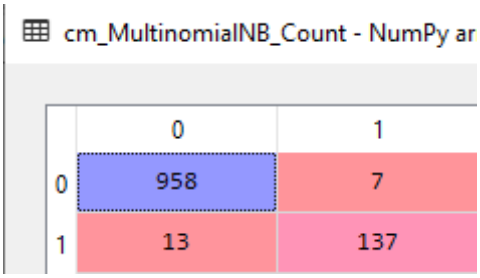

Fig. 16. Multinomial Naive Bayes Confusion Matrix

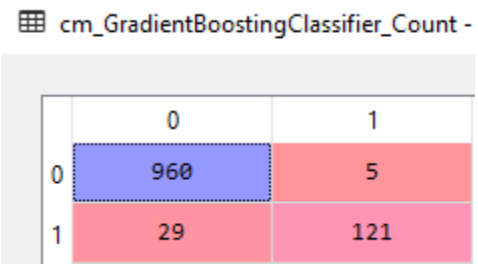

Fig. 17. Gradient Boosting Confusion Matrix

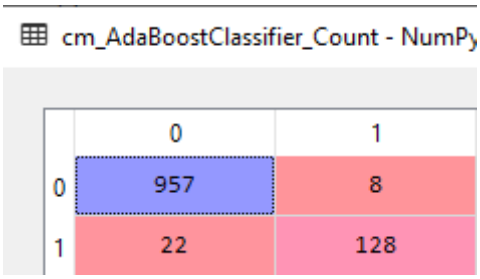

Fig. 18. Ada Boosting Confusion Matrix

Performance analysis is done by analyzing the performance metrics like Accuracy, FScore, Precision and Recall is shown in Table 1 – Table 2 anf Fig 19 – Fig 22.

Table. 1. Estimation of Classifier Parameters

Classifier Methods	Precision	Recall
KNN classifier	0.92	0.92
Random Forest classifier	0.98	0.98
Linear SVM	0.98	0.98
Ada Boost classifier	0.97	0.97
Kernel SVM	0.89	0.89
Logistic Regression	0.98	0.98
Gaussian Naive Bayes	0.91	0.83
Decision Tree classifier	0.97	0.97
Extra Tree classifier	0.96	0.96
Gradient Boosting	0.97	0.97
Multinomial Naive Bayes	0.98	0.98

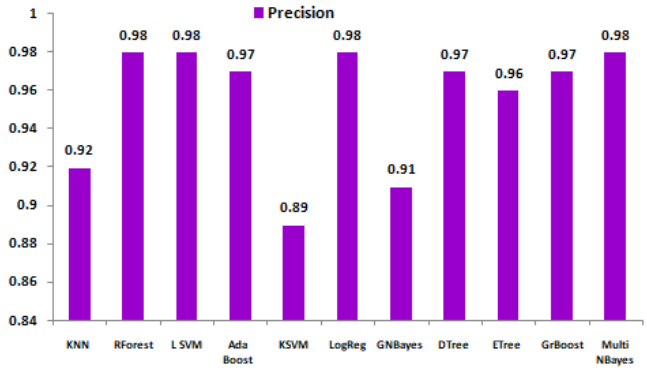

Fig. 19. Precision Analysis

Fig. 20. Recall Analysis

Table. 2. Accuracy Estimation of Classifier Parameters

Classifier Methods	Fscore	Accuracy (%)
KNN classifier	0.91	92.28
Random Forest classifier	0.98	97.84
Linear SVM	0.98	97.58
Ada Boost classifier	0.97	97.30
Kernel SVM	0.89	89.32
Logistic Regression	0.98	97.57
Gaussian Naive Bayes	0.85	82.95
Decision Tree classifier	0.97	96.95
Extra Tree classifier	0.96	96.23
Gradient Boosting	0.97	96.94
Multinomial Naive Bayes	0.98	98.20

Retrieval Number: D7287118419/2019©BEIESP
DOI:10.35940/ijrte.D7287.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

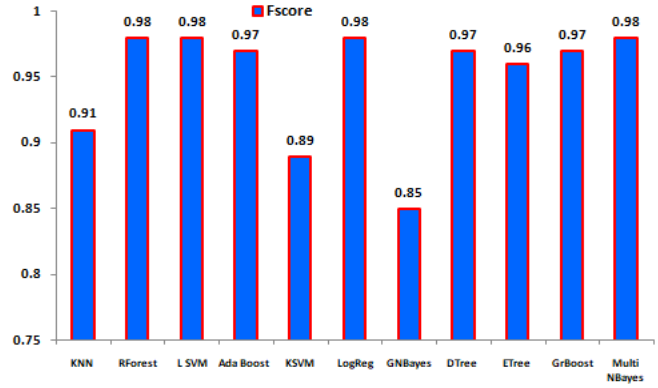

Fig. 21. FScore Analysis

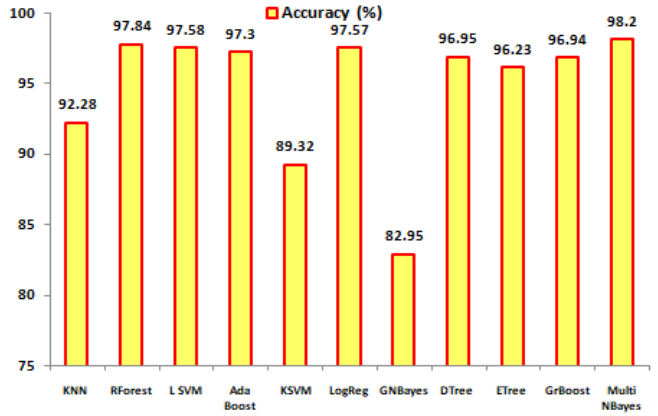

Fig. 22. Accuracy Analysis

V. CONCLUSION

This paper attempts to predict the spam and ham message detection in SMS Spam Message detection dataset from the KAGGLE machine learning repository. The prediction of fraud transaction is done by using machine learning claasification algorithms. Then the top most high correlated word features are extracted from the SMS Spam Message detection dataset using Count vectorizer. Experimental Results shows that the Multinomial Naive Bayes classifier have achieved the effective prediction with the precision of 0.98, recall of 0.98, FScore of 0.98 , and Accuracy of 98.20%.

REFERENCES

1. Joseph Janison, "Applying Machine Learning to Predict Davidson College's Admissions Yield", proceedings of the ACM SIGCSE Technical Symposium., 2017.
2. William Eberle, Douglas Talbert, Erik Simpson, Larry Roberts, and Alexis pope, " Using Machine Learning and Predictive Modeling to Assess Admission Policies and Standards", Proceedings of the 9th Annual National Symposium., 2013.
3. 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.
4. 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.
5. 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.
6. 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
7. 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
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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).
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. 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.
19. 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.
20. 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.
21. 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.