Stratification of Spam and Ham Short Message Service using Machine Learning Hash Vectorization

M. Shyamala Devi, M. Nizar Ahamed, S. Aarif Ahamed

Abstract*: With the increase in the usage of mobile technology, the rate of information is duplicated as a huge volume. Due to the volume duplication of message, the identification of spam messages leads to challenging task. The growth of mobile usage leads to instant communication only through messages. This drastically leads to hackers and unauthorized users to the spread and misuse of sending spam messages. The identification of spam messages is a research oriented problem for the mobile service providers in order to raise the number of customers and to retain them. With this overview, this paper focuses on identifying and prediction of spam and ham messages. The SMS Spam Message Detection dataset from KAGGLE machine learning Repository is used for prediction analysis. The identification of spam and ham messages is done in the following ways. Firstly, the levels of spread of target variable namely spam or ham is identified and they are depicted as a graph. Secondly, the essential tokens that are responsible for the spam and ham messages are identified and they are found by using the hashing Vectorizer and it is portrayed in the form of spam and Ham messages word cloud. Thirdly, the hash vectorized SMS Spam Message detection dataset is fitted to various classifiers like Ada Boost Classifier, Extra Tree classifier, KNN classifier, Random Forest classifier, Linear SVM classifier, Kernel SVM classifier, Logistic Regression classifier, Gaussian Naive Bayes classifier, Decision Tree classifier, Gradient Boosting classifier and Multinomial Naive Bayes classifier. The evaluation of the classifier models are done by analyzing the Performance analysis metrics like Accuracy, Recall, FScore, Precision and Recall. The implementation is done by python in Anaconda Spyder Navigator. Experimental Results shows that the Linear Support Vector Machine classifier have achieved the effective performance indicators with the precision of 0.98, recall of 0.98, FScore of 0.98 , and Accuracy of 98.71%.

***Index Terms*: Machine Learning, Hashing, Vectorizer, Spam, Ham and Classifier**

I. INTRODUCTION

In machine learning, the usage of mobile technology leads to the large volume of duplicated information and the usage of instant short message service. The unauthorized users or hackers use the technological advancements to share the spam messages to the entire users of the various mobile service providers. Now, it is the challenging task for the mobile service providers to satisfy the customers in order to retain in their service. The customer gets disappointed if they continuously receive the spam messages from the unauthorized unknown persons. Nowadays the hackers use the spam messages to access the customer's personal information from their mobile node.

Generally, users store all their personal secured password details in their mobile. So it is the responsibility of the user to protect the password details without making others to access. But the technological hacking advancement has made to access the personal secured information of the mobile users by just viewing the spam messages spent by the unauthorized users. The spam message program is written in such a way that some background software is made to execute when the mobile users view the spam message. If the users continuously receive the spam message, then they may switch their service provider. So it may affect the business of the service provider too. This leads to the usage of machine learning algorithms to predict the spam messages and to discard or unsubscribe those unwanted messages based on the early predictions.

The paper is organized by discussing the existing details in the related work with Section 2 followed by the proposed work in the Section 3. Implementation and the performance execution analysis are discussed in Section 4 followed by the conclusion of the paper in Section 5.

II. RELATED WORK

A. Literature Survey

The graph data mining techniques are used to differentiate the spammers and nonspammers in the messages that are assumed to be indispensible. The prediction of spammers and nonspammers in the messages are done without directly checking the original message contents [1]. The short message service spam message is the unnecessary annoyance to the mobile users. The service providers business has the problem of losing the customers subscription towards their service. A broad review of spam filtering and detection techniques is critically reviewed and analyzed [2]. The spam mobile message is harmful to the mobile node as the technology grows. The spam message detection is done based on the frequency of particular content in the mobile message. The spam message prediction is done based on the machine learning algorithms by using biasian and spam filters [3].

Revised Manuscript Received on January 15, 2020

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

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

S. Arif Ahamed, Assistant Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

The volume of spam email is drastically increasing as the mobile users are in frequent use. Bayesian filters are used to stop the spam messages by entering the mobile node. The Bayesian filters are used to block the spam message and emails from our mobile [4]. The statistical data compression models can be used to filter the spam messages. Spam The spam filtering model is designed as a probabilistic text classifiers based on character-level in the message and the occurrence of binary sequences in the message content. The empirical evaluation outperforms the spam filters [5].

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 [6] – [30].

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 levels of spread of target variable namely spam or ham is identified and they are depicted as a graph.

(ii) Secondly, the essential tokens that are responsible for the spam and ham messages are identified and they are found by using the hashing Vectorizer and it is portrayed in the form of spam and Ham messages word cloud.

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

(iv) Fourth, the evaluation of the classifier models are done by analyzing the Performance analysis metrics like Accuracy, Recall, FScore, Precision and Recall.

A. System Architecture

The overall architecture model of this paper is shown in Fig. 1

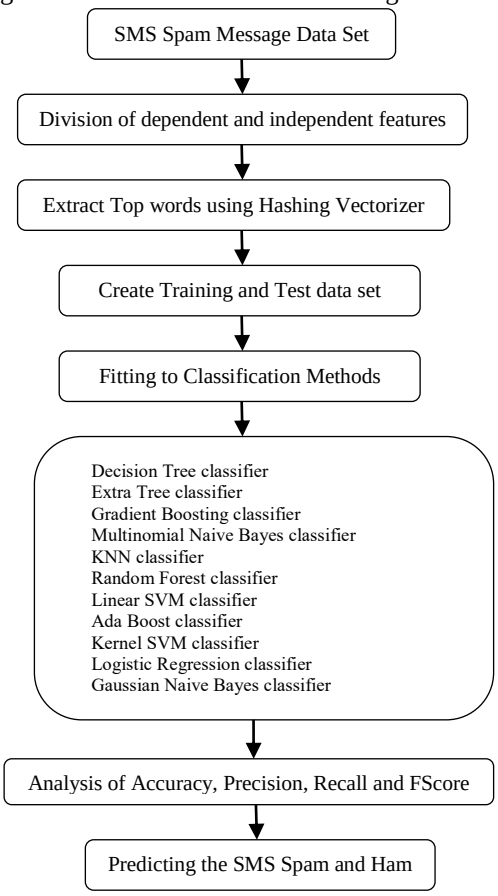

Fig. 1 System Architecture

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 / Ham Type (Target- Dependent Attribute)

SMS Spam Message detection Data Set is executed to analyze the target distribution of Spam/Ham class type and is shown in Fig. 2.

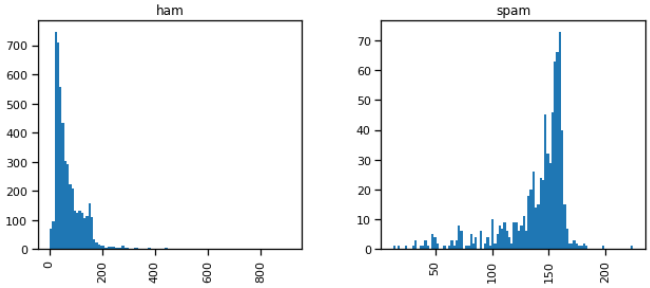

Fig. 2 Dataset Spam/Ham Target division

B. Prediction of SMS Spam Message detection

SMS Spam Message detection Data Set is executed to categorize the level and the number of target Spam and Ham messages in the dataset and is shown in Fig. 3.

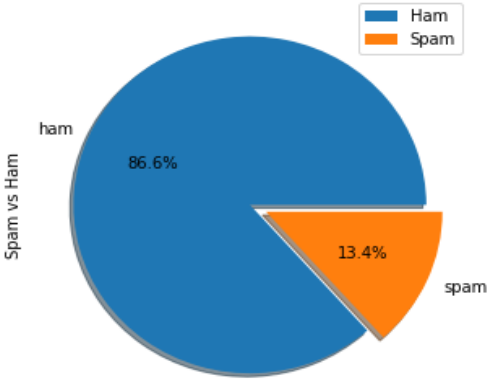

Fig. 3 Spam and Ham Message level in the Dataset

The top words of the spam and ham messages in the dataset are identified from the dataset using Hashing Vectorizer and it is shown in Fig. 4 – Fig. 5.

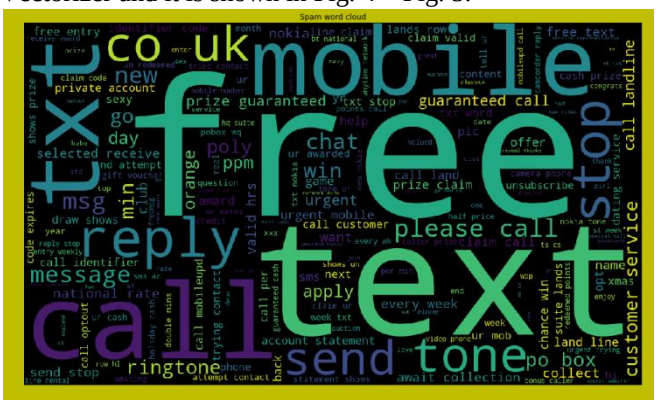

Fig. 4. Spam Word Cloud

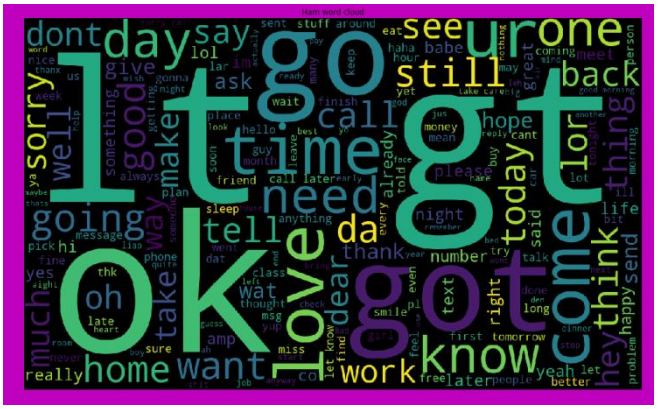

Fig. 5. Ham Word Cloud

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

```
In [9]: Data['Category'].value_counts()
Out[9]:
ham     4825
spam     747
Name: Category, dtype: int64
```

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

df1 - DataFrame

Index	Words
0	go
1	jurong
2	point
3	crazy
4	available
5	bugis
6	n
7	great
8	world
9	la
10	e
11	buffet
12	cine
13	got

Fig. 7 Spam and Ham message Words

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

cm_logreg_HashingfVectorizer - NumP

	0	1
0	964	1
1	35	115

Fig. 8. Logistic Regression Classifier Matrix

cm_knn_Count_HashingVectorizer - Nu

	0	1
0	965	0
1	67	83

Fig. 9. KNN Classifier Matrix

cm_Linearsvm_HashingVectorizer - Nur

	0	1
0	962	3
1	21	129

Fig. 10. Linear SVM Classifier Matrix

cm_kernelsvm_HashingVectorizer - Nur

	0	1
0	965	0
1	150	0

Fig. 11. Kernel SVM Classifier Matrix

cm_NB_GaussianNB_HashingfVectorizer

	0	1
0	865	100
1	15	135

Fig. 12. Gaussian Naive Bayes Classifier Matrix

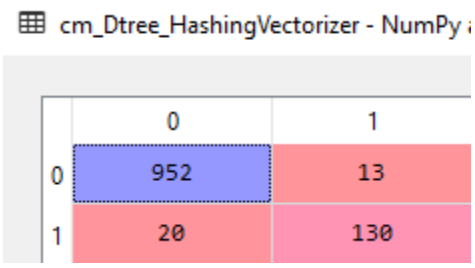

Fig. 13. Decision Tree Classifier Matrix

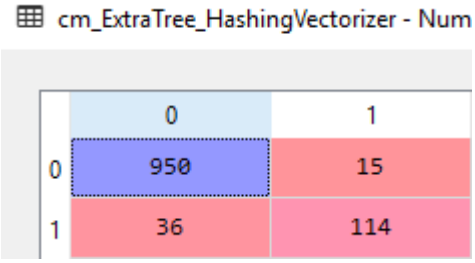

Fig. 14. Extra Tree Classifier Matrix

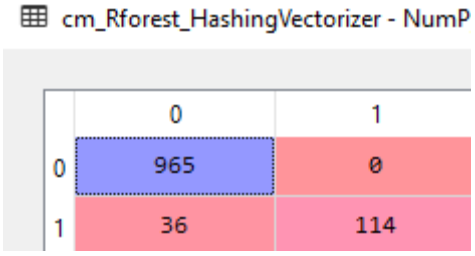

Fig. 15. Random Forest Classifier Matrix

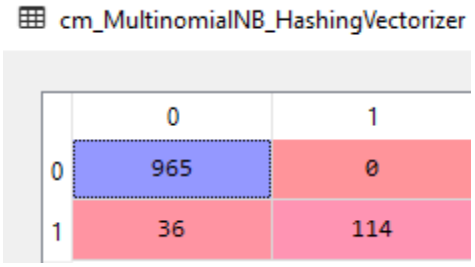

Fig. 16. Multinomial Naive Bayes Classifier Matrix

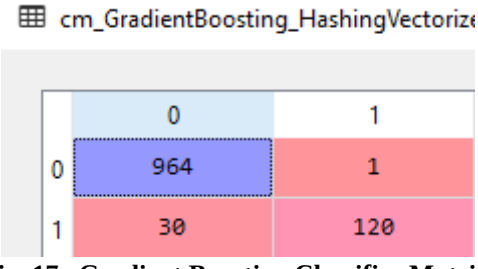

Fig. 17. Gradient Boosting Classifier Matrix

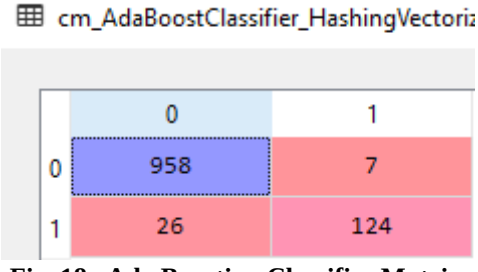

Fig. 18. Ada Boosting Classifier Matrix

Execution performance analysis is done by analyzing the performance metrics like Accuracy, FScore, Precision and Recall is shown in Table 1 – Table 2.

Table. 1. Precision and Recall Estimation of Classifier

Classifier Methods	Precision	Recall
Decision Tree classifier	0.96	0.96
Extra Tree classifier	0.95	0.95
KNN classifier	0.93	0.94
Random Forest classifier	0.97	0.97
Linear SVM	0.98	0.98
Kernel SVM	0.75	0.87
Logistic Regression	0.97	0.96
Gaussian Naive Bayes	0.93	0.90
Multinomial Naive Bayes	0.97	0.97
Gradient Boosting	0.96	0.97
Ada Boost classifier	0.97	0.97

Table. 2. Accuracy Estimation of Classifier Parameters

Classifier Methods	FScore	Accuracy (%)
Decision Tree classifier	0.97	97.14
Extra Tree classifier	0.95	95.42
KNN classifier	0.93	93.99
Random Forest classifier	0.97	96.71
Linear SVM	0.98	98.71
Kernel SVM	0.80	86.54
Logistic Regression	0.95	95.67
Gaussian Naive Bayes	0.91	89.68
Multinomial Naive Bayes	0.97	96.77
Gradient Boosting	0.97	97.21
Ada Boost classifier	0.97	97.04

V. CONCLUSION

This paper identifies the occurrence and the analysis of predicting the spam and ham message in SMS Spam Message detection dataset from the KAGGLE machine learningrepository. The prediction of spam and ham message is done by using machine learning hash vectorizer and classification algorithms. Experimental Results shows that the Linear Support Vector Machine classifier have achieved the effective performance indicators with the precision of 0.98, recall of 0.98, FScore of 0.98 , and Accuracy of 98.71%.

REFERENCES

1. Qian Xu, Evan Wei Xiang, Qiang Yang, Jiachun Du and Jieping Zhong, "SMS Spam Detection using Non-Content Features", IEEE Intelligent Systems, vol. 27, no. 6, 2012, pp. 44-51.
2. Shafii Muhammad Abdulhamid, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi, Abdul-Salaam, Adamu I. Abubakar, and Tutut Herawan, "A Review on Mobile SMS Spam Filtering Techniques", IEEE Access, 99, 2017.
3. Ashok Koujalagi, "Mobile SMS Spam Recognition Using Machine Learning Techniques with the help of Biasian and Spam Filters", International Journal of Computer Sciences and Engineering, vol. 7, no. 4, 2017, pp. 540-542.
4. Jose Maria Gomez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sanz and Francisco Carrero Garcia, "Content based SMS spam filtering", : Proceedings of the 2006 ACM Symposium on Document Engineering, Amsterdam, The Netherlands, October 2006.
5. Bogdan Filipic, Gordon V. Cormack, Thomas R. Lynam and Blaz Zupan, "Spam Filtering Using Statistical Data Compression Models",The Journal of Machine Learning Research archive, vol. Volume 7, 2006, pp. 2673-2698.
6. 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.
7. 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.
8. 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.
9. 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
10. 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
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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).
20. 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).
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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
30. 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.