International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
IJCAI 2021; 2(2): 41-46
Received: 22-05-2021
Accepted: 24-06-2021

Dhanunjaya B
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Pavan Kumar Reddy B
Assistant Professor, Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Corresponding Author:
Dhanunjaya B
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

PCA based email spam detection utilizing machine learning approaches

Dhanunjaya B and Pavan Kumar Reddy B

DOI: https://doi.org/10.33545/27076571.2021.v2.i2a.37

Abstract

Email is perhaps the most secure mechanism for online correspondence and moving information or messages through the web. A congesting expansion in notoriety, the quantity of spontaneous information has likewise expanded quickly. To sifting information, various methodologies exist which naturally identify and eliminate these unsound messages. Spam messages impact the associations monetarily as well as bother the individual email client. Presently a day's Machine learning frameworks are utilized to thus channel the spam email in an extraordinarily powerful rate. The reason for this investigation is to recognize significant diminished information highlights in building IDS that is computationally proficient and viable. In this paper, we propose a structure for highlight decrease utilizing head part investigation (PCA) to wipe out unessential highlights, choosing pertinent and non-related highlights without influencing the data contained in the first information and afterward utilizing choice tree and SVM calculations to group email information. The point is to lessen a few highlights of information by PCA and afterward assemble a forecast of characterization model by choice tree and SVM to acquire pertinent highlights and to improve the precision of choice tree and SVM. Observational outcomes show that chose decreased traits give better execution to configuration spam mail discovery framework that is productive and powerful for Email spam sifting.

Keywords: spam, PCA, SVM, decision tree and ML

1. Introduction

Email framework is quite possibly the best and normally utilized wellsprings of correspondence. The explanation of the fame of email framework lies in its financially savvy and quicker correspondence nature. Shockingly, email framework is getting undermined by spam messages. Spam messages are the excluded messages sent by some undesirable clients otherwise called spammers with the intention of bringing in cash [5]. The email clients invest a large portion of their significant energy in arranging these spam sends. Different duplicates of same message are sent ordinarily which influence an association monetarily as well as disturbs the accepting client. Spam messages are meddling the client's messages as well as delivering huge measure of undesirable information and in this way influencing the organization's ability and use. In this paper, a Spam Mail Detection (SMD) framework is proposed which will characterize email information into spam and ham messages. The interaction of spam sifting centers around three fundamental levels: the email address, subject and substance of the message [6].

For spam email, clients are dealing with a few issues like maltreatment of traffic, limit the extra room, computational force, become a boundary for tracking down the extra email, burn through client's time and furthermore danger for client security [7]. In this way, turning out to be email safer and viable, suitable Email sifting is fundamental.

All sends have a typical design for example subject of the email and the body of the email. An average spam mail can be ordered by sifting its substance. The cycle of spam mail location depends with the understanding that the substance of the spam mail is not the same as the real or ham mail. For instance, words identified with the promotion of any item, support of administrations, dating related substance and so on the cycle of spam email location can be comprehensively classified into two methodologies: information designing and AI approach [5]. Information designing is an organization-based methodology in which IP (web convention) address, network address alongside some arrangements of characterized rules is considered for the email grouping. The methodology has shown promising outcomes yet it is very tedious. The support and assignment of refreshing guidelines isn't helpful for all clients. Then again, AI approach doesn't include any arrangement of rules and is proficient than information designing methodology [3]. The arrangement calculation orders the email dependent on the substance and different traits.

The destinations of this paper are:

- To study different highlights of information dataset
- To apply the Principal Component Analysis (PCA) for lessening the quantity of characteristics
- To group information utilizing Decision Tree and SVM

2. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality decrease strategy used to change high-dimensional datasets into a dataset with less factors, where the arrangement of coming about factors clarifies the greatest difference inside the dataset [1, 8]. PCA is utilized before unaided and managed AI steps to lessen the quantity of highlights utilized in the investigation, in this way decreasing the probability of mistake. The PCA will recognize designs in the informational collection, and discovers their similitudes and contrasts between each credit. It goes about as an incredible model to dissect the information. This depends on the thought that countless information sources, while expanding the computational burden, don't really add to improving the viability of the classification [9, 10].

The essential segments are a straight line, and the primary head segment holds the most change in the information. Each ensuing head part is symmetrical to the last and has a lesser difference. PCA is transcendently utilized as a dimensionality decrease procedure in areas like facial acknowledgment, PC vision and picture pressure. It is additionally utilized for discovering designs in information of high measurement in the field of account, information mining, bioinformatics, brain science, and so forth the primary direct method for dimensionality decrease, head segment examination, plays out a straight planning of the information to a lower-dimensional space so that the difference of the information in the low-dimensional portrayal is expanded. The PCA will recognize designs in the informational index, and discovers their likenesses and contrasts between each credit [8]. It goes about as an incredible model to break down the information. PCA will take all the first preparing set factors and break down them in a way to make another arrangement of factors with high clarified fluctuation.

3. Proposed Methodology

In this proposed system, include decrease strategy utilizing PCA is led as an underlying advance towards lessening the quantity of qualities without losing the primary reason and target data from the first information. The following stage is building up an indicator with an improved precision to group informational index.

3.1 Support Vector Machine (SVM)

SVM is an AI calculation that can be displayed for both relapse and grouping issues yet it is significantly utilized for characterization of a double class issue [9]. The fundamental thought from SVM was to verifiably plan the preparation informational collection into a high dimensional component space. At the point when a marked preparing information is given as an info, the model gives an ideal hyperplane as a yield, which classifies the examples. It is not difficult to keep a straight hyperplane between two classes. The essential thought from SVM was to verifiably plan the preparation informational collection into a high dimensional component space. The idea of the SVM strategy is to make an ideal hyperplane what isolates information into two classes [4]. The ideal hyperplane is a territory what parts the information into classes and it is found opposite to the nearest design. Examples are spots that depict an informational collection. To get the ideal hyperplane, we need to track down the most extreme edge. Edge is a reach between the hyperplane with the nearest design for each class, while support vector is the closest example to the ideal hyperplane.

3.2 Decision Tree

A Decision Tree is a tree-like chart comprising of inside hubs which address a test on a property and branches which mean the result of the test and leaf hubs which connote a class name. The order rules are shaped by the way chose from the root hub to the leaf. To isolate each info information, first the root hub is picked as it is the most noticeable trait to isolate the information. The tree is built by distinguishing credits and their related qualities which will be utilized to investigate the information at each moderate hub of the tree [3, 4]. After the tree is framed, it can prefigure recently coming information by navigating, beginning from a root hub to the leaf hub visiting every one of the interior hubs in the way relying on the test states of the traits at every hub. The fundamental benefit of utilizing choice trees rather than other order procedures is that they give a rich arrangement of decides that are straightforward.

4. Experimental Results

The trial was led with double center 2.20 GHz with 4.00 Go of memory on windows stage, and we executed the calculation utilizing weka [3]. Weka represents Waikato Environment for Knowledge Analysis. Weka is made by analysts at the University of Waikato in New Zealand. Weka was first executed in quite a while current structure in 1997. The product is written in the Java language and contains a GUI for associating with information documents. For working of weka, we needn't bother with the profound information on information digging for which weka an exceptionally well-known information is mining instrument. weka additionally gives the graphical UI of the client and gives numerous offices. Weka is a cutting-edge office for creating AI (ML) procedures and their application to certifiable information mining issues. The information document ordinarily utilized by weka is in ARFF record design. ARFF represents Attribute Relation File Format, which comprises of extraordinary labels to show separating in the information record. Weka executes calculations for information pre-preparing, grouping, relapse and bunching and affiliation rules. It likewise incorporates representation devices. It has a bunch of boards, every one of which can be utilized to play out a specific undertaking. The new AI plans can likewise be created with this bundle.

The target of this segment is to assess our proposed calculation regarding precision, number of chosen highlights, and learning exactness on chose highlights. In this test, the PCA is utilized to decrease irrelevant highlights. The proposed PCA based choice tree and SVM has been tried E-mail spam base of UCI dataset [11]. This dataset comprises of 4,601 number of examples and 57 number of traits. The last section of this information base means whether the email is viewed as spam (1) or not (0). There are 1813 spam occasions and 2788 non-spam cases. The vast majority of properties show that how continuous for a word or character to happen in a given email. Also, the excess ascribes measure the length of groupings of successive capital letters. It contains a bunch of spam and non-spam classes. A large portion of the qualities show whether a specific word or character is regularly happening in the email or not. We haphazardly pick a preparation set (70%) and a testing set (30%) of three informational collections. To approve the expectation consequences of the proposed strategy k-overlay hybrid approval is utilized. Measurable rundown of information as demonstrated in the figure-1.

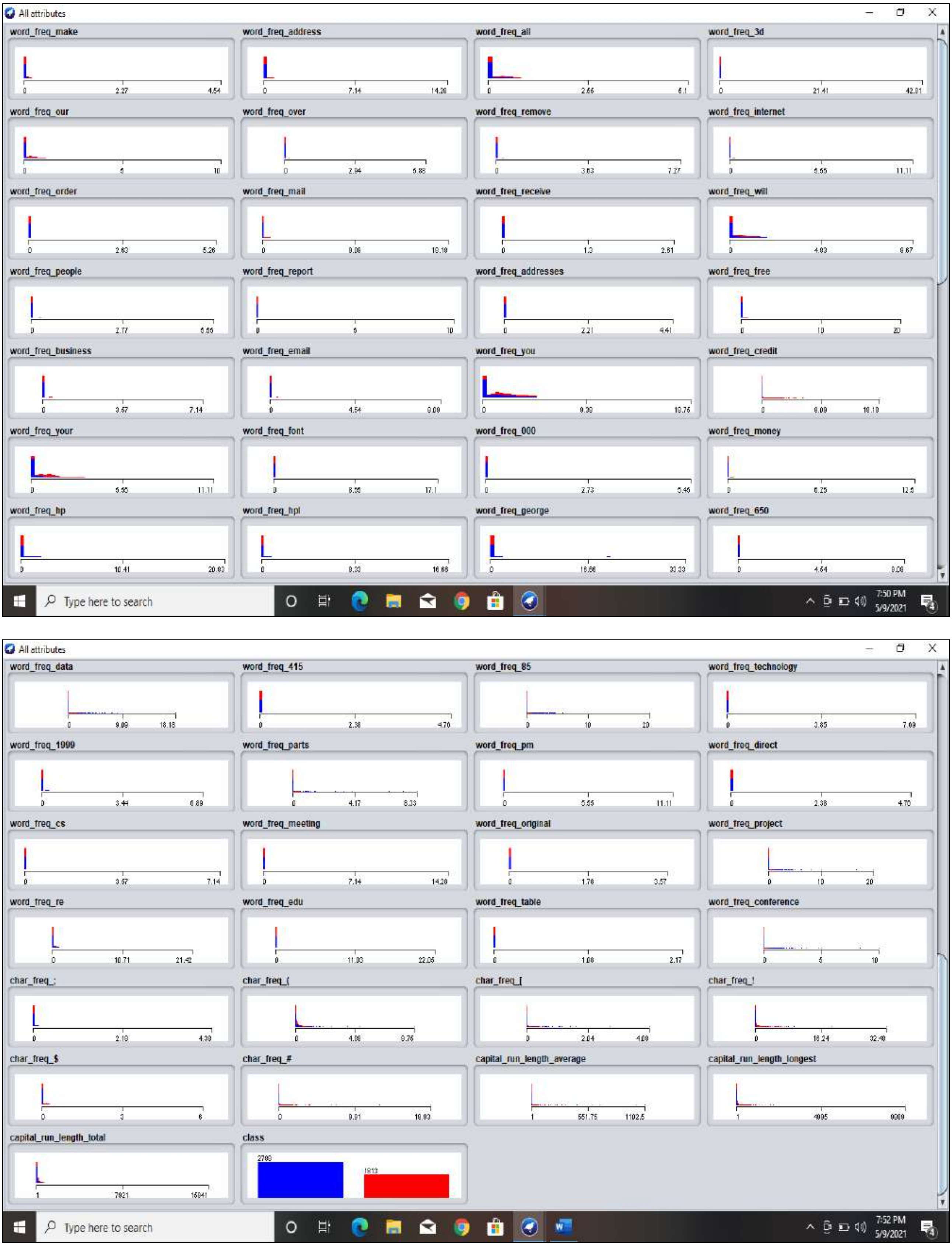
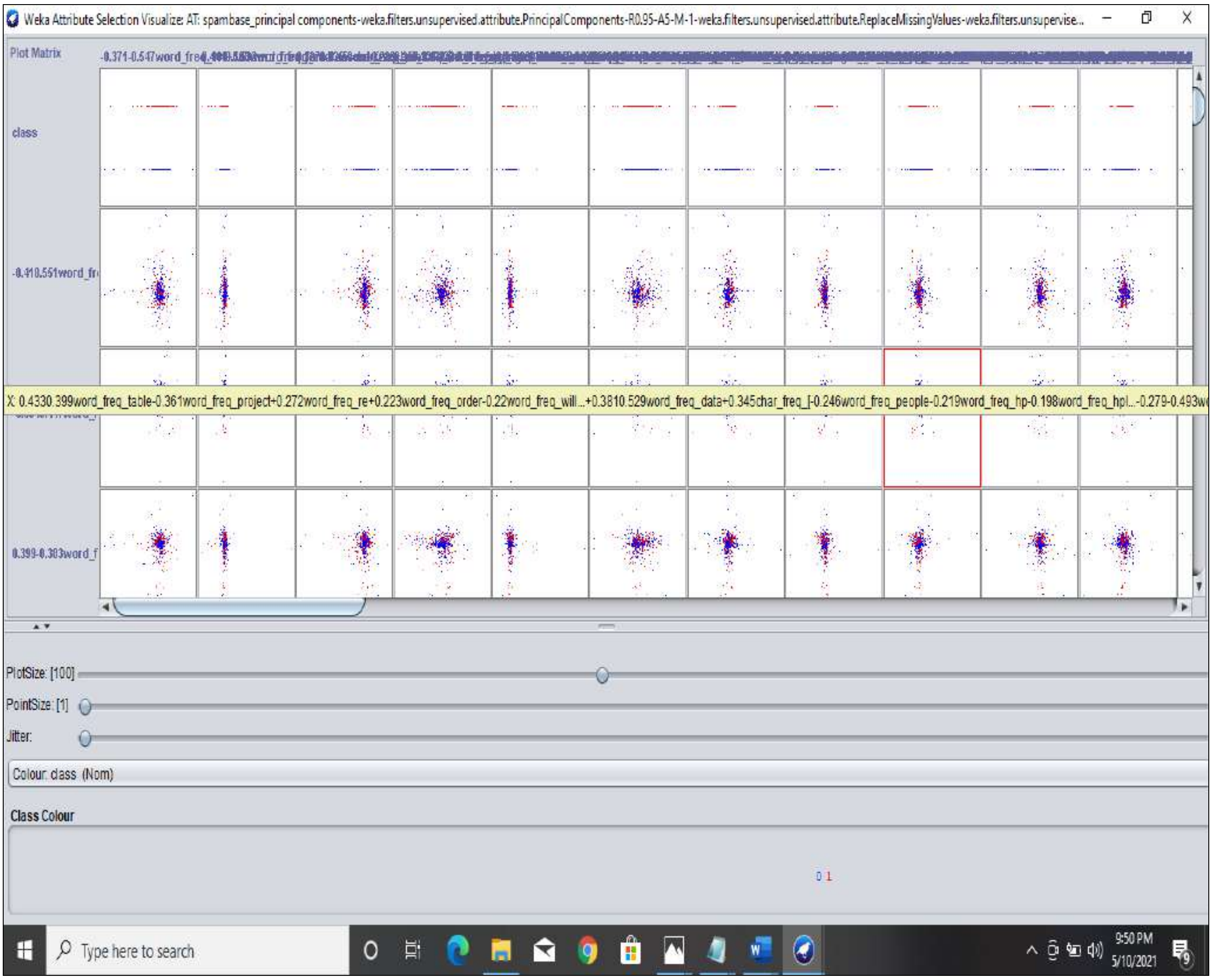

All attributes
word_freq_make 0 2.27 4.54
word_freq_address 0 7.14 14.28
word_freq_all 0 2.55 5.1
word_freq_3d 0 21.41 42.81
word_freq_our 0 5 10
word_freq_over 0 2.94 5.88
word_freq_remove 0 3.63 7.27
word_freq_internet 0 5.55 11.11
word_freq_order 0 2.63 5.26
word_freq_mail 0 9.09 18.18
word_freq_receive 0 1.3 2.61
word_freq_will 0 4.83 9.67
word_freq_people 0 2.77 5.55
word_freq_report 0 5 10
word_freq_addresses 0 2.21 4.41
word_freq_free 0 10 20
word_freq_business 0 3.57 7.14
word_freq_email 0 4.54 9.09
word_freq_you 0 9.38 18.75
word_freq_credit 0 9.09 18.18
word_freq_your 0 5.55 11.11
word_freq_font 0 8.55 17.1
word_freq_000 0 2.73 5.46
word_freq_money 0 6.25 12.5
word_freq_hp 0 10.41 20.83
word_freq_hpl 0 8.33 16.66
word_freq_george 0 16.66 33.33
word_freq_650 0 4.54 9.09
Type here to search
7:50 PM
5/9/2021
All attributes
word_freq_data 0 9.09 18.18
word_freq_415 0 2.38 4.76
word_freq_85 0 10 20
word_freq_technology 0 3.85 7.69
word_freq_1999 0 3.44 6.89
word_freq_parts 0 4.17 8.33
word_freq_pm 0 5.55 11.11
word_freq_direct 0 2.38 4.76
word_freq_cs 0 3.57 7.14
word_freq_meeting 0 7.14 14.28
word_freq_original 0 1.78 3.57
word_freq_project 0 10 20
word_freq_re 0 10.71 21.42
word_freq_edu 0 11.03 22.05
word_freq_table 0 1.08 2.17
word_freq_conference 0 5 10
char_freq_; 0 2.19 4.38
char_freq_(0 4.38 9.76
char_freq_[0 2.04 4.08
char_freq_! 0 16.24 32.48
char_freq_$ 0 3 6
char_freq_# 0 9.01 19.03
capital_run_length_average 1 551.75 1102.5
capital_run_length_longest 1 4995 9989
capital_run_length_total 1 7921 15841
class 2788 1813
Type here to search
7:52 PM
5/9/2021

Fig 1: Statistical representation of data

4.1 Result and Discussion

The dataset is huge and high-dimensional in nature. Consequently, both datasets go through dimensionality decrease utilizing PCA. Figure 2 shows the fluctuation assessment of the 22 head segments got for both datasets as they contain similar 49 highlights. The important parts having a combined variety of over 80% are held and the leftover ones are disposed of based on total variety which is over 90% while the excess segments are disposed of due to the leftover 20% variety which is immaterial. The parts with higher difference will be held in the boundary choice outcomes after dimensionality decrease utilizing PCA.



Weka Attribute Selection Visualize: AT: spambase_principal components-weka.filters.unsupervised.attribute.PrincipalComponents-R0.95-A5-M-1-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervise...

X: 0.4330.399word_freq_table-0.361word_freq_project+0.272word_freq_re+0.223word_freq_order-0.22word_freq_will...+0.3810.529word_freq_data+0.345char_freq_[-0.246word_freq_people-0.219word_freq_hp-0.198word_freq_hpl...-0.279-0.493w

Fig 2: Variance of Principal components

From the Spam base information 49 trait we have sifted to 22 element vectors by utilizing PCA strategy to get an ideal determination from complete dataset for preparing just as for testing tests. Table-1 shows the test precision that accomplished by utilizing the two calculations for the full measurement information and furthermore after the element decrease with PCA strategy. The fundamental 22 highlights eliminated by PCA add to essentially less fluctuation and thus the leftover 27 credits chose.

Table 1: Performance of Classifiers with PCA and without PCA

Algorithm	Accuracy	Precision	Recall
Decision Tree without PCA	90.93	91	90.9
Decision Tree with PCA	92.19	92.2	92.2
SVM without PCA	92.97	93	93
SVM with PCA	95.5	95.5	95.5

Figure-2 shows the test accuracy on class E-mail Spam classification that compared with 49 features and with the reduced (27) set of features by using PCA technique.

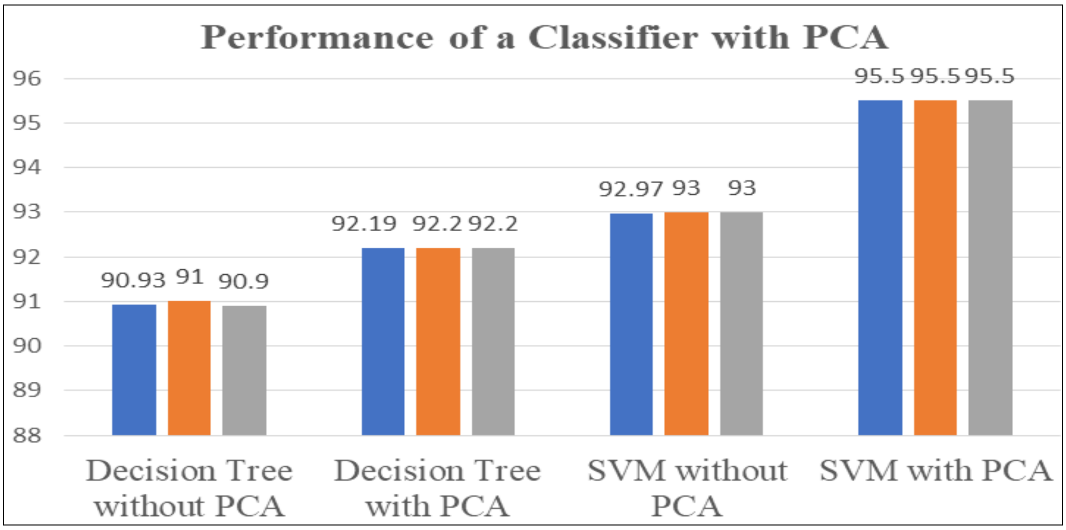
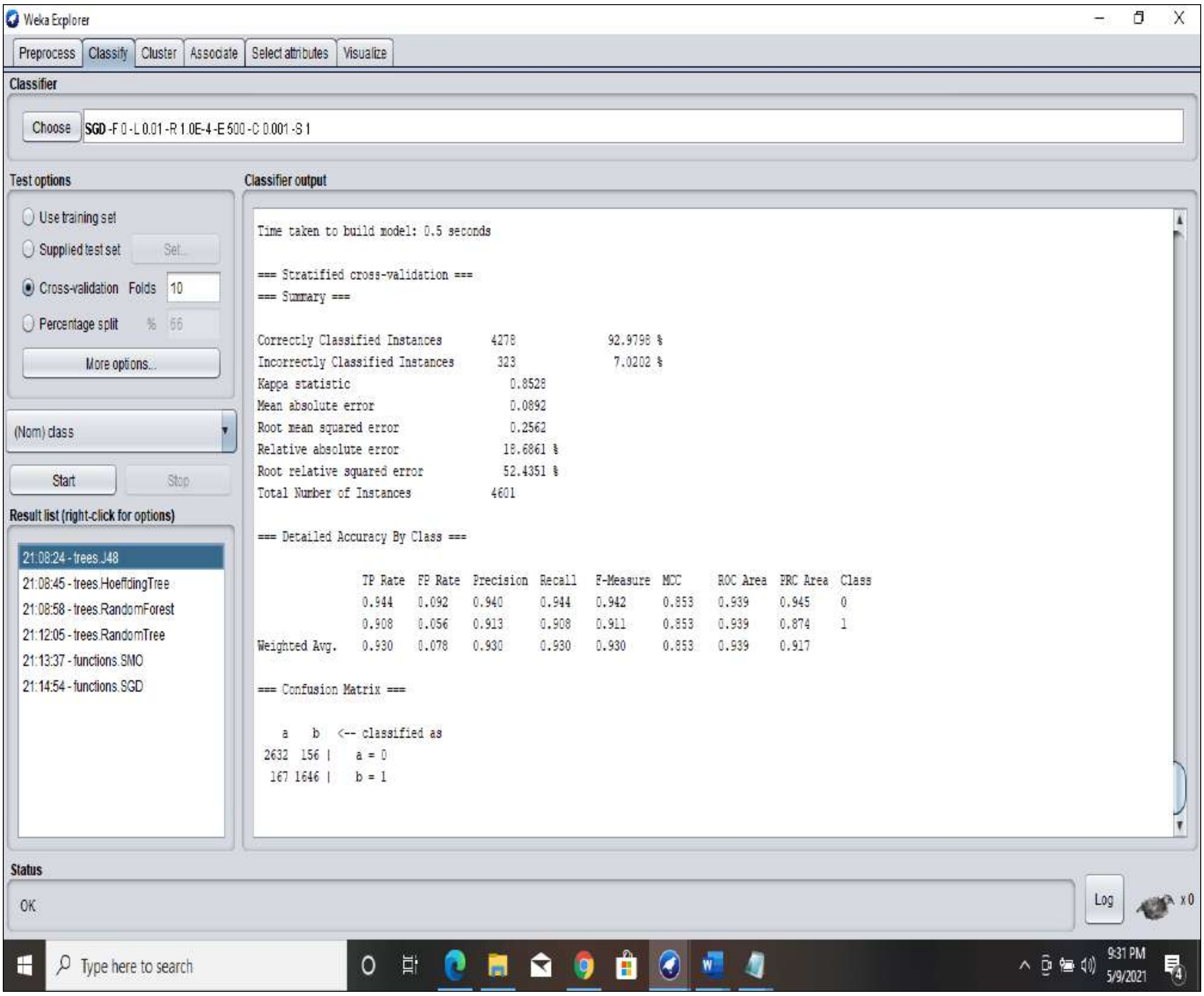


Fig 3: Performance of ML Algorithms

We see in the figure-3, the presentation of model Decision tree without PCA dependent on exactness, exactness, and review esteems are 90.93%, 91% and 90.0% separately, while the exhibition of choice tree with PCA dependent on exactness, accuracy and review esteems are 92.19%, 92.2% and 92.2%. Here the choice tree with PCA calculation shows the most noteworthy exactness contrasted and choice tree without PCA. We likewise notice execution of SVM without PCA dependent on exactness, accuracy, and review esteems are 92.97%, 93% and 93% separately, while the exhibition of SVM with PCA dependent on dependent on precision, exactness and review esteems are 95.5%, 95.5% and 95.5%. In this way, the general presentation has accomplished SVM with PCA are better performance when compared to decision tree.



SGD -F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1

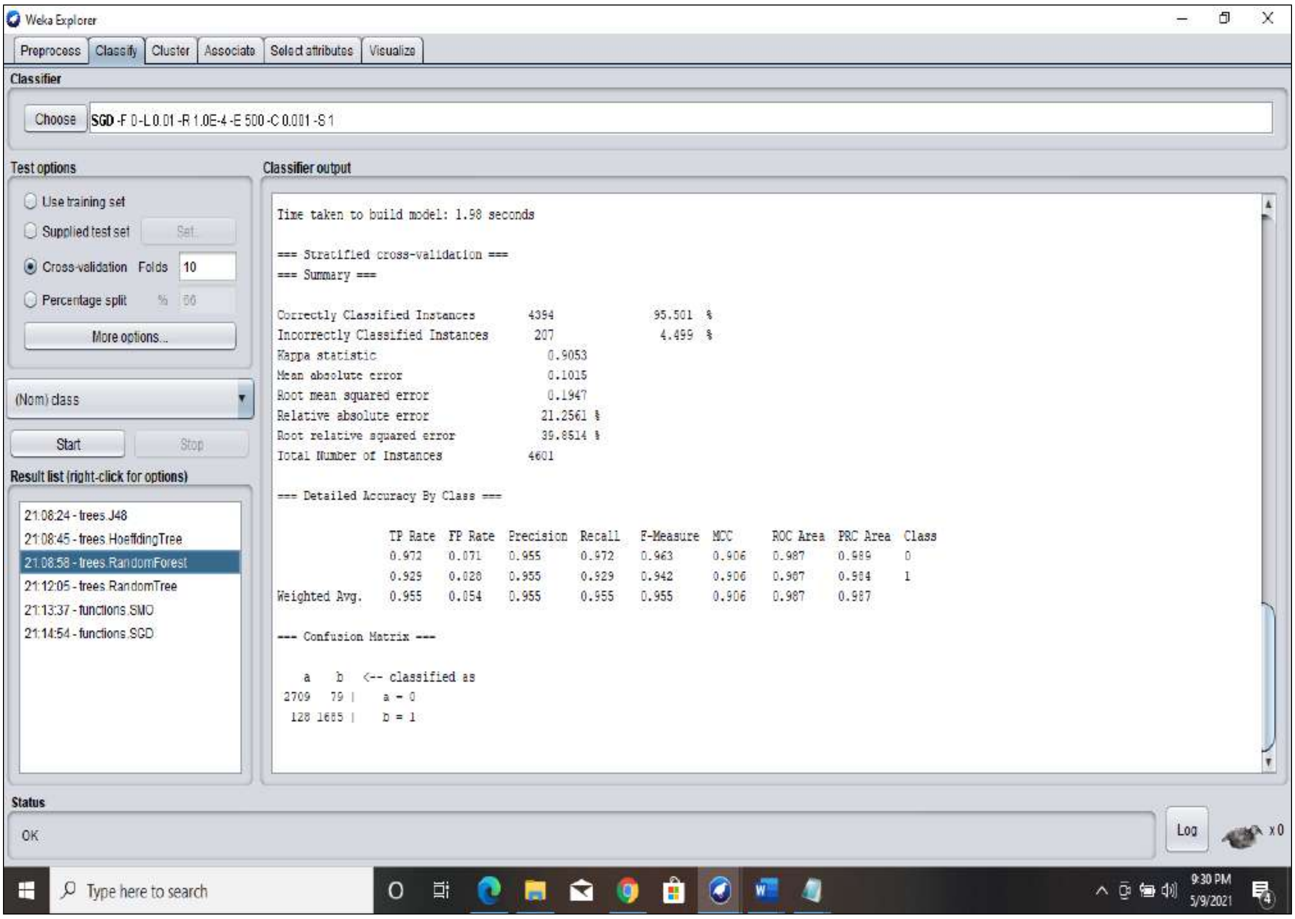
Folds 10

○ Percentage split % 66

21:08:24 - trees.J48
21:08:45 - trees.HoeffdingTree
21:08:58 - trees.RandomForest
21:12:05 - trees.RandomTree
21:13:37 - functions.SMO
21:14:54 - functions.SGD

```
Time taken to build model: 0.5 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4278               92.9798 %
Incorrectly Classified Instances       323                7.0202 %
Kappa statistic                          0.8528
Mean absolute error                      0.0892
Root mean squared error                  0.2562
Relative absolute error                 18.6861 %
Root relative squared error             52.4351 %
Total Number of Instances             4601

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944    0.092    0.940      0.944    0.942      0.853    0.939     0.945     0
                 0.908    0.056    0.913      0.908    0.911      0.853    0.939     0.874     1
Weighted Avg.    0.930    0.078    0.930      0.930    0.930      0.853    0.939     0.917

=== Confusion Matrix ===

    a    b   <-- classified as
 2632  156 |    a = 0
  167 1646 |    b = 1
```

OK



SGD -F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1

● Cross-validation Folds 10

(Nom) class

```
21:08:24 - trees.J48
21:08:45 - trees.HoeffdingTree
21:08:58 - trees.RandomForest
21:12:05 - trees.RandomTree
21:13:37 - functions.SMO
21:14:54 - functions.SGD
```

```
Time taken to build model: 1.98 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4394               95.501  %
Incorrectly Classified Instances       207                4.499  %
Kappa statistic                          0.9053
Mean absolute error                      0.1015
Root mean squared error                  0.1947
Relative absolute error                 21.2561 %
Root relative squared error             39.8514 %
Total Number of Instances             4601

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.071    0.955      0.972    0.963      0.906    0.987     0.989     0
                 0.929    0.028    0.955      0.929    0.942      0.906    0.987     0.984     1
Weighted Avg.    0.955    0.054    0.955      0.955    0.955      0.906    0.987     0.987

=== Confusion Matrix ===

    a    b   <-- classified as
 2709   79 |    a = 0
  128 1685 |    b = 1
```

OK

5. Conclusion

This paper proposes an E-mail Spam location model coordinating PCA and grouping calculations (SVM and choice tree). Dimensionality decrease utilizing PCA eliminates uproarious characteristics and holds the ideal quality subset. SVM and Decision tree develop arrangement models dependent on preparing information acquired from PCA. The result of the PCA is to project a component space onto a more modest subspace that addresses information by lessening the elements of highlight space. This diminishes computational expenses and the blunder of boundary assessment. It was reasoned that the proposed PCA based SVM calculation performs better execution. After the investigation it ought to foresee that, SVM with PCA calculation is the best calculation in AI procedure and can better order of email spam.

6. References

1. Ravi Kumar G, Nagamani K, Anjan Babu G. A Framework of Dimensionality Reduction Utilizing PCA for Neural Network Prediction, Lecture Notes on Data Engineering and Communications Technologies, ISBN 978-981-15-0977-3. Springer Nature Singapore Pte Ltd 2020;37:173-180.
2. Ravi Kumar G, Venkata Sheshanna Kongara, Dr. Ramachandra GA. An Efficient Ensemble Based Classification Techniques for Medical Diagnosis, International Journal of Latest Technology in Engineering, Management and Applied Sciences 2013;2(8):5-9. ISSN-2278-2540.
3. Witten H, Frank E. Data mining: practical machine learning tools and techniques with Java implementations, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc 2000.
4. Han J, Kamber M. Data Mining concepts and Techniques, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann 2006.
5. DíAz NP, OrdáS DR, Riverola FF, MéNdez JR. SDAI: An integral evaluation methodology for content-based spam filtering models, Expert Systems with Applications 2012;39(16):12487-12500.
6. Guzella TS, Caminhas WM. A review of machine learning approaches to spam filtering, Expert Systems with Applications 2009;36(7):10206-10222.
7. Ma W, Tran D, Sharma D. A novel spam email detection system based on negative selection, In: Proc. of Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT'09, Seoul, Korea 2009, 987-992.
8. Lei X. A novel feature extraction method assembled with PCA and ICA for network intrusion detection, Comput. Sci. Technol. Appl. IFCSTA 2003;3:31-34.
9. Li Y *et al*. An efficient intrusion detection system based on support vector machines and gradually feature removal method, Expert Systems with Applications 2012;39:424-430. http://dx.doi.org/10.1016/j.eswa.2011.07.032
10. Pang Y, Yuan Y, Li X. Effective feature extraction in high dimensional space. IEEE Trans. Syst 2008.
11. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/.