Logistic Regression Based Classification of Spam and Non-Spam Emails

Shahbaz Ahmad Khanday[1], Suraiya Parveen[2]
{shahbazshaban10@gmail.com[1], husainsuraiya@gmail.com[2]}

Jamia Hamdard University New Delhi India[1,2]

Abstract. An email client receives emails from different websites, portals and domains, which can be an advertisement. Receiving a bulk amount of emails can cause serious damages like suspension of a particular email id. Mostly an email client gets exposed to the number of malicious receipts by registering an email account to a web portal, which in turn sends a bulk amount of emails. One of the solutions to escape from spam emails is to develop a decision based system which can classify the spam and non-spam emails. This can be achieved using different machine learning and deep learning and deep learning algorithms to classify the spam and non-spam emails by accessing the received emails of an email client. The machine learning approaches and mechanisms like SVM, naive Bayesian classifier, artificial neural networks and random forests can be of important help to determine spam emails. After classifying a spam email source a user can navigate, block and report the source of the spam email generator like spam-bots.

Keywords: machine learning, decision tree, support vector machine (SVM), logistic regression, artificial neural networks, naive Bayesian classifier and spam-bots.

1 Introduction

A common person can receive a huge amount of emails in a day. The email user can receive emails from different sources related to the different day to day activities like social networking, files and sharing, online shopping, e billing, e commerce and applications etc. One should be able to differentiate between important and useful emails over spam or junk emails. Once a user gets exposed to the spam and malicious sources he will receive a large amount of emails from various unknown sources. Therefore it becomes a hectic and time consuming task for an email user to make a selection and difference of all the received emails, which may contain an important piece of data or information. The condition becomes very risky when an email client is trapped into a malicious act and then the security and privacy of a system could be breached. The email user could be trapped into a phishing act initiated by the cyber criminals. It is very hard to recover from such situations and most of the times an email user gets attracted to the spam emails and respond to them. In most of the cases the blocking and reporting of these spam email sources become useless, as the senders change their location continuously. One of the alternatives can be tracking those particular IP addresses from where an email user receives these spam emails, but the task becomes harder when the number of IP addresses are many but not fewer. And the major part is when the senders change their locations and targets. One of the

ICIDSSD 2020, February 27-28, New Delhi, India

DOI 10.4108/eai.27-2-2020.2303291

solutions to the email spamming is to access an email id with the number of senders, classify and categorize the received emails into spam and non-spam emails. The classification of received emails can be done by using appropriate and approximate machine learning approaches and some autonomous algorithms like support vector machine, artificial neural networks, logistic regression and naive Bayesian classifier. All of machine learning approaches use decision tree based modeling and testing on a given input data set and produce the result in few classified groups [14]. The decision tree performs testing and examination on data at its nodes. The resultant branch of the tree is the outcome of the test performed at the node of the decision tree [32]. The spam and non-spam emails can be further classified into many more sub groups. A particular email sender which is repeated in the inbox of the email client can be captured and examined. Which in case can be a spam mail source and user can be provided options to report or block such sources.

1.1 Machine learning

Machine learning (ML) is a category of algorithm that allows software applications to became more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available [24][22]. The processes involved in machine learning are similar to that of data mining and predictive modelling. Both require searching through data to look for patterns and adjusting program actions accordingly. Many people are familiar with machine learning from shopping on the internet and being served ads related to their purchase [22]. This happens because recommendation engines use machine learning to personalize online ad delivery in almost real time. Beyond personalized marketing, other common machine learning use cases include fraud detection, spam filtering, network security threat detection, predictive maintenance and building news feeds.[35]

Deep learning

Deep learning is the subset of machine learning. Deep learning is an aspect of artificial intelligence (AI) that is concerned with emulating the learning approach that human beings use to gain certain types of knowledge. At its simplest, deep learning can be thought as a way to automate predictive analytics. Deep learning algorithms are very different when compared with the machine learning algorithms.[42] Deep learning algorithms and approaches are used to predict, analyse and process large datasets. Models and algorithms like deep belief networks, convolutional neural networks and recurrent neural networks fall in the category of deep leaning which are used for processing, predicting and analysing domains like image recognition, computer vision, speech recognition, brief designs, natural language processing, recognition of data from social networks and bioinformatics etc.[14][21]

The deep learning models are so autonomous which learn from the events by automatically changing the parameters and weights on the edges of the networks for optimal and accurate solutions. The network models also support the feed forward and feedback approaches to train the models [14]. The different types of network architectures popular in the field of the deep learning are:-

1. Deep belief networks.
2. Convolutional neural networks.
3. Recurrent neural networks.

1.2 Email Bombing and Spamming:-

Email Spamming is like sending an email to large number of emails id on a single click. These activities are performed mainly for advertisements of products or services provided by a company. Spam sources and frauds use automated software to collect emails available on websites, forms, chat rooms and the malicious and spam emails are sent to the email clients.[33][35]

Email bombing is an activity of sending a large and enormous numbers of emails to a target email, resulting in damage of email id.

The mails received by the email client are called junk emails. In this case a email client receives a spam emails from different email ids. Storage of these emails is a burden for email server. The email server keeps notifying the client by sends alarms and notifications to the client. The email service provider may block that particular client from services. The conviction of spam emails has been rapidly growing and gaining popularity since the early years of 1990s and it is the common problem faced by almost every email user. The sources of the spam emails and the receipts obtain the email addresses from the spam-bots,[33] which are autonomous and automated programs and applications that seek and search the internet looking for email addresses, which could be used as a target. Spammers use spam-bots to generate and create a distribution list containing emails. A spammer or a spam-bot typically keeps sending emails to millions of email clients using the email addresses, with the expectation that only fewer or a small number of clients will respond, read and interact with the message.

1.3 Types of Spams

Spam emails comes in the different kind of forms, the most popular being to promote and advertise outright scams, phishing or newly approached publications and business schemes. Spam email is typically the one used to promote access to inexpensive pharmaceutical drugs, aesthetics and athleticism, online courses, job portals, job opportunities and online bidding and gambling.[33] Spam is commonly used to conduct email fraud like phishing. The other concerns of using the spam emails are credit card and funding fraud, where a client is brought into the act so that the banking and credit card details and credentials could be retrieved from the client. The other forms of the spam emails are the contents like banners for promoting events, short videos and other web contents like webpages, forums, documents, images, xml and html files.

2 Related Work

This section discusses some of the prominent work done in the field of spam detection. Many scholars have realized the need for new methods of detecting spams since social networking is rising and does not offer any mechanism to provide secure identification system.

Nikihila et. al. [1] observes the techniques for reducing the logistic loss function in the spam filtering problem and carries out performance analysis of different techniques. The goal of this paper is to identify if the email is spam or not and recognizes logistic regression as one of the best technique to categorize an email as spam or not spam. Three different type of algorithm for minimization of logistic regression are studied and implemented-Stochastic Gradient Descent Algorithm, Regular Batch Gradient Descent Algorithm, and Regularized Gradient Descent Algorithm. The paper determines that it is unclear to optimally control the weight vector in Stochastic Gradient Descent algorithm which works on simulated annealing technique, whereas performance was upgraded on the test set in normal gradient descent, as it stopped overfitting in the training data.

Qingha et. al. [2] carries out survey on regularly used approaches to thwart e-cheating, and demonstrate how biometrics can be used for this purpose. The author puts forward a new method to observe student activities by using their IP addresses and timestamps to contribute in observing potential cheat behavior. The outcomes show that the proposed method is effective at recognizing student collision during exam.

Moein sarvi et. al. [3] runs a fuzzy expert system that is used for detection of spams. The proposed model uses several email features to prepare a fuzzy model which then results in an expert system followed by defuzzification process. The developed system was tested with sets of 1000, 2000, 3000 4000 messages and the best outcomes were achieved with the set of 3000 messages. The system is measured using Recall and Precision criteria and the best results obtained were 97.4% and 99.3% respectively.

Shadi khawandi et. al. [4] share their concern over image spam detection since it has been serious issue over the years and numerous solutions have been provided by different vendors. This paper focuses on the process used for preventing spams while explaining the available solutions for handling spam and image based spam. The paper concludes that the available anti-spam methods are not sufficient as most of the mail servers count on the blacklists whereas others depend on filters that might convey high false positive rate.

Idris et. al. [5] familiarizes us with an email detection system that is considered as an enhancement in the negative selection algorithm (NSA). Particle swarm optimization (PSO) was applied to recover the random detector generation in the negative selection algorithm (NSA). The hybrid which is achieved by combining NSA–PSO practices a local outlier factor in terms of the fitness function designed for the detector generation. The detector generation method is then concluded as the estimated spam coverage is reached. After this step, the enhancement of the uniqueness between the non-spam and spam detectors is carried out by a distance measure and a threshold value. The analysis shows that the accuracy of the proposed hybrid NSA–PSO model is better than the accuracy of the standard NSA model. The proposed model can be used to distinguish between spam and non-spam in a network.

Authors [6] introduce a spam detection technique. This technique uses text clustering method. This method gave an efficient model by realizing contents of various email and detect spam. This technique observes clusters with the help of spherical k means algorithm and clusters into two groups, that is spam and non-spam. Centroid vectors are obtained for extracting the description of clusters. For each vector in centroid, the label, whether the email is spam or not is assigned by checking the number of spam email in the cluster. Finally, the label of the most appropriate cluster is allocated to the new mail. The upshots show accuracy in this model is somewhat near to support vector machine (SVM).

Priyanka Sao et. al. [6] compares the performance of Naïve Bayes classifier with support vector machine. The aim of this paper is spam classification and author suggests naïve Bayesian classifier to be one of the simplest and efficient methods for the classification of spam. Results show that naïve Bayesian classifier has more accuracy than support vector machine since the error rate is very low in Naïve Bayes classifier.

Kamoru et. al. [7] targets to investigate available research works in spam detection approaches, the process which is being followed in these methods, and other mitigation systems. Many anti-spam strategies are surveyed for email and social networking in this paper. The author enlightens the importance of working on spam detection for the betterment of the world. This study reveals the new issues and challenges which needs to be addressed and is a big challenge for research.

Kamoru et, al. sights approach carries out a study on different algorithms for the purpose of spam detection. [8] These algorithms are studied under two groups, that is, content based filtering and rule based filtering. Several techniques on content based filtering have been calculated and investigated in this paper. It is concluded that rule based filtering is most efficient method to create spam filter since it reduces the filtering time.

In this section [9], a synopsis of the spam filtering is examined and the methods researchers use for evaluation and comparison purposes of these different methods are analyzed. This paper gives the gist of spam detection algorithms which come under the class of content based filtering. The results show that Bayesian classifier correctly classifies at the accuracy of 96.5%, followed by Chi square test, which gives the result at 92%, whereas KNN classifier has the accuracy of 89%.

Krasser S et. al. [10] evaluates classification performance results for C4.5 decision tree and support vector machine for detecting image spam. The analyses conclude that feature extraction is considered to pose a very low computational load and the classification is partial towards a low false positive rate. About 60% of spam images can be eliminated using the techniques with a low false rate of 0.5%. Therefore the model proposed in this paper serves as an efficient first tier framework to detect large amount of spam images without doing expensive calculations.

Malarvizhi et. al, have analyzed dataset using TANAGRA data mining tool and explores the efficient classifier for detecting email spams. [11] Feature construction and selection is done to extract relevant features which is followed by classification algorithms and cross validation is done over this dataset. The paper approves Random Forest tree classification as the best classifier since it produces more than 99% accuracy in spam detection. This Random Forest classifier is tested with test dataset and gives accurate results than other classifiers for this spam dataset.

Krasser et, al. provides the report which classifies the tweets into spam and non-spam using machine learning techniques. The results show that Naïve Bayes gives better results than Support Vector Machine (SVM). The data set is decomposed into training and testing , 70% for training sand 30% for testing. When the data set is applied in SVM , it is first trained and then tested and got 76% accuracy. And then same data set is tested with Naïve Bayes where it is trained and tested and got 92%. The results may vary based on the feature selection. [12]

The section defines the summary of diagnoses and predicting the factors of multiple diseases in aged and elderly persons. The survey includes the summary of disease called pneumonia using multiple logistic regression algorithms. The authors has also analyzed the comparison of pneumonia disease targets differentiating the risk factor of losing lives among younger and elderly people. [13]

The goal is to differentiate the multiple machine learning and deep leaning functions which could be used for development and improvement in cyber security. [14] Using various machine learning and deep learning autonomous and smart algorithms could be developed for the betterment of cyber activities and security. The author in the survey report points out the similarities and dissimilarities between various machine learning and other learning algorithms which namely are support vector machine, K-nearest neighbor, decision trees, fuzzy systems and Logistic regression etc. The survey includes an email data set for analysis and decision tree based approach.

The author in the paper titled “K-means Clustering Algorithm” defines each and every step along with required functions to perform the clustering on a given dataset.[15]

The author of the book “Machine Learning & Pattern recognition Series” is Stephen Marsland in which he discusses the applications of machine learning in pattern recognition.

The authors of paper 18 and 20 [18][20] performed the implementation of the machine learning algorithm support vector machine in the newly emerged domains of pattern recognition.

The papers which are published in the favor of the decision support system are paper 19, paper 27 and paper 30 [19][27][30]. The researchers in these papers had defined the decision tree as one of the key algorithms for the autonomous systems to co-operate in the modern world. Also the definition and accuracy of decision tree ID3 had been crucial step in their research.

S. Nasser, R. Alkhaldi, and G. Vert et.al. and F. Sets have implement their scholarly thoughts and examinations in different datasets by means of fuzzy sets and fuzzy systems.

M. E. Tipping and C. M. Bishop et.al. had gone through the fact of using decision trees as one the common principles of machine learning algorithms as a probabilistic classifier.

3. Methodology

3.1 Dataset Description

We use data set tilted "spam-base dataset" for classification and modeling of decision tree, which is freely available at UCI machine learning database [acchives.ics.uci.edu] This set of data includes the words and the characters in the words defining the frequency of and the occurrence of the email receipts. The algorithms works on the dynamics and procedure of word counting, which basically could recognize the special words occurring the number of times in an email inbox. Therefore by using a special function the similar and repeating words could be distinguished. The describe function can be used in python jupyter to demonstrate the word and character frequency after reading the contents of the dataset.

3.2 Logistic Regression

Logistic regression is one of the most likely and appropriate algorithm used for classification of datasets. In case of classifying a dataset named as spam base the logistic regression is the most versatile decision based approach for detecting spam emails out of a dataset. Logistic regression performs some basic test on the given distribution of data which involves finding and calculating some statistical domains like mean and standard deviation.[16] It also produces results of the operations like word and character count, max and min operations. After producing and provisioning the statistical and count tests the logistic regression algorithm fetches the outcome of these tests and tends to inter-relate the outcomes. A table below demonstrates the

definition of one of the basic functions used in logistic regression. Logistic regression is a basic mechanism in statistical analysis that attempts to predict a data value based some prior distributions and observations. A logistic regression algorithm forms the relationship between one dependent variable and one or more dependent variables.
Demonstration and comparison of various machine learning and deep learning algorithms for the classification of spam and non-spam emails. The different machine learning and deep learning algorithms could be compared and distinguished by means of definition and accuracy for classifying a dataset into number of categories or variables.

Table 3.1.2. Comparison of different algorithms.

Algorithm used for classification	Definition of the Algorithm	Accuracy of classifying objects
Chi square function	Algorithms come under the class of content based filtering.	92%
Naïve Bayes classifier	Based upon the combination of multiple algorithms.	75-92%
Fuzzy systems	The system is measured using Recall and Precision criteria.	92%
Decision tree C4.5	Extension of previous decision trees.(ID3)	60%
SVM	The result of SVM is a hyper-plane.	76%
KNN	Finds out the nearest neighbor by voting.	89%
Logistic regression	Defines the interdependency between dependent and independent variables.	80%

3.3 Proposed Model

The project can be designed after following some crucial steps, which includes:-

1. Dataset containing spam and non-spam emails.
2. Classification by using the suitable ML or DL algorithm.
3. IP or email addresses reading and extraction.

4. Testing.
5. Blocking and reporting.

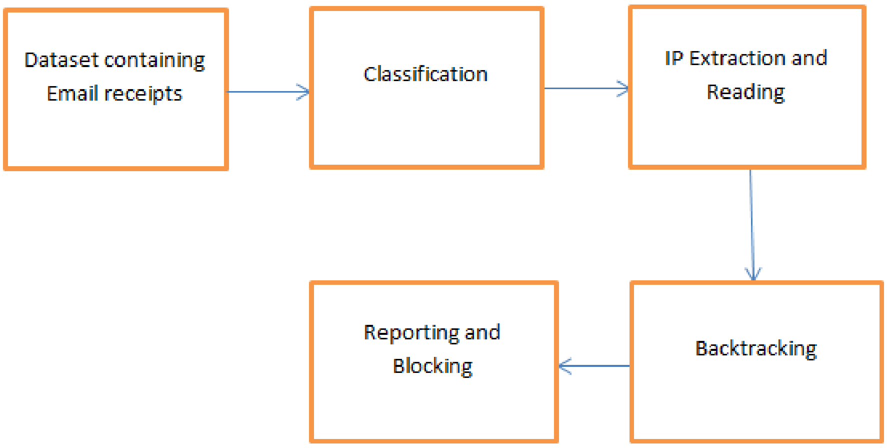


Figure 3.3. Block Diagram of email classification model

3.3.1 Classification Model

The proposed model defines the basic structure of the reading and extracting the IP addresses of the emails received by an email client. The model tests the sources of the spam emails once the classification of spam and non-spam emails is performed using any suitable machine learning or deep learning classification techniques.

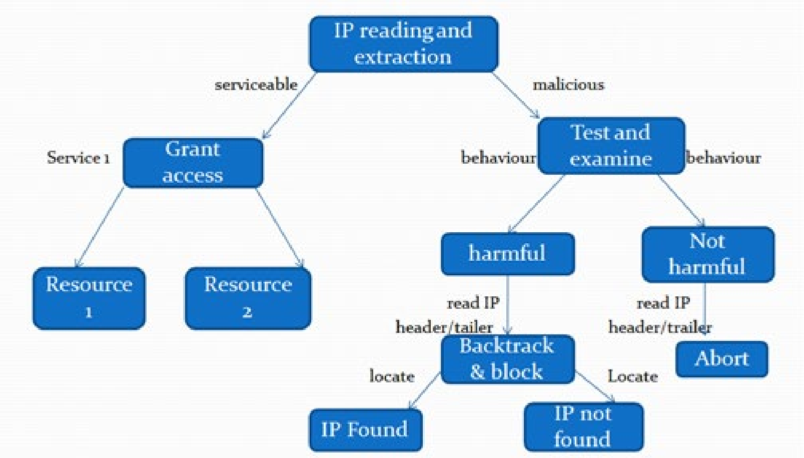


Figure 3.3.2 A decision tree based systematic model for reading and blocking spam email sources

The following model demonstrates the phases which could be followed to read, test and block a spam email source. At first the IP address of the receipt is read and extracted by checking the catalog and the information of the received spam email.. Then the access could be granted to the non-malicious email and for extraction of the information contained by the particular email. The next phase of the model is to test and examine the occurrence as well as the information contained by the email id. The proposed model of the classification application is quite similar to the decision tree based systematic model with few alterations during the examination of the email receipts. The model begins with the process by analyzing the email receipts with unique email addresses. The algorithm examines the repeating emails in the reception list and predicts the repeating emails as the spam emails. In most of the cases the spam email source keeps sending malicious emails repeatedly. The operations performed on the email addresses are the character count/character frequency or word count/word frequency to outcome one or more spam email senders.

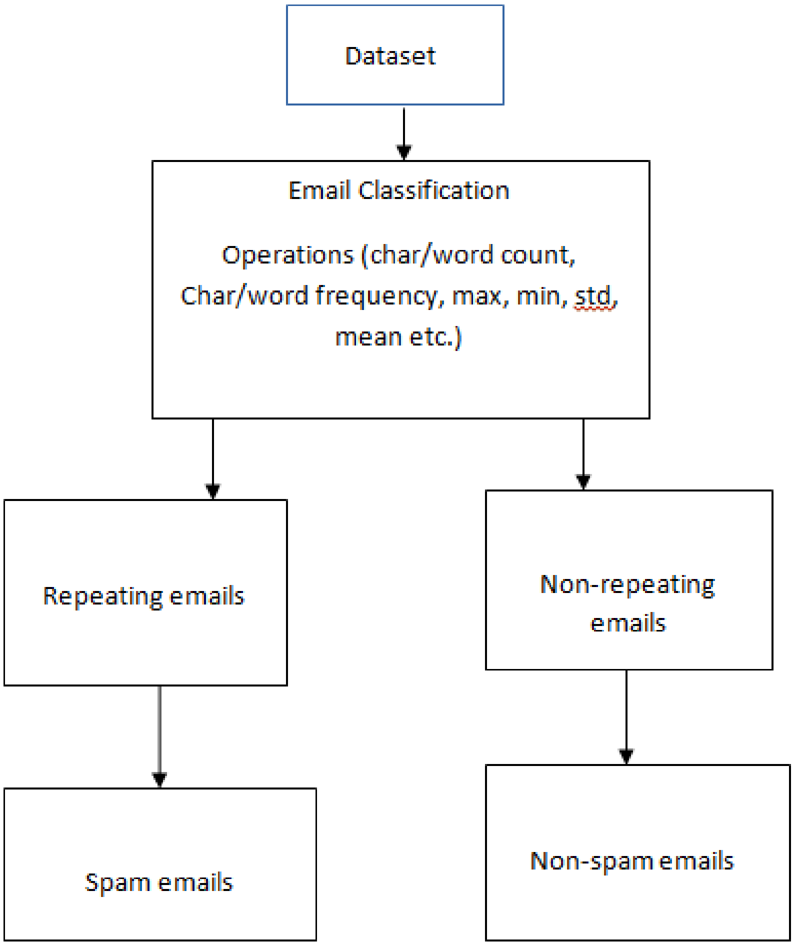


Figure 3.3.3 Flowchart depicting the procedure of the algorithm

4. Results

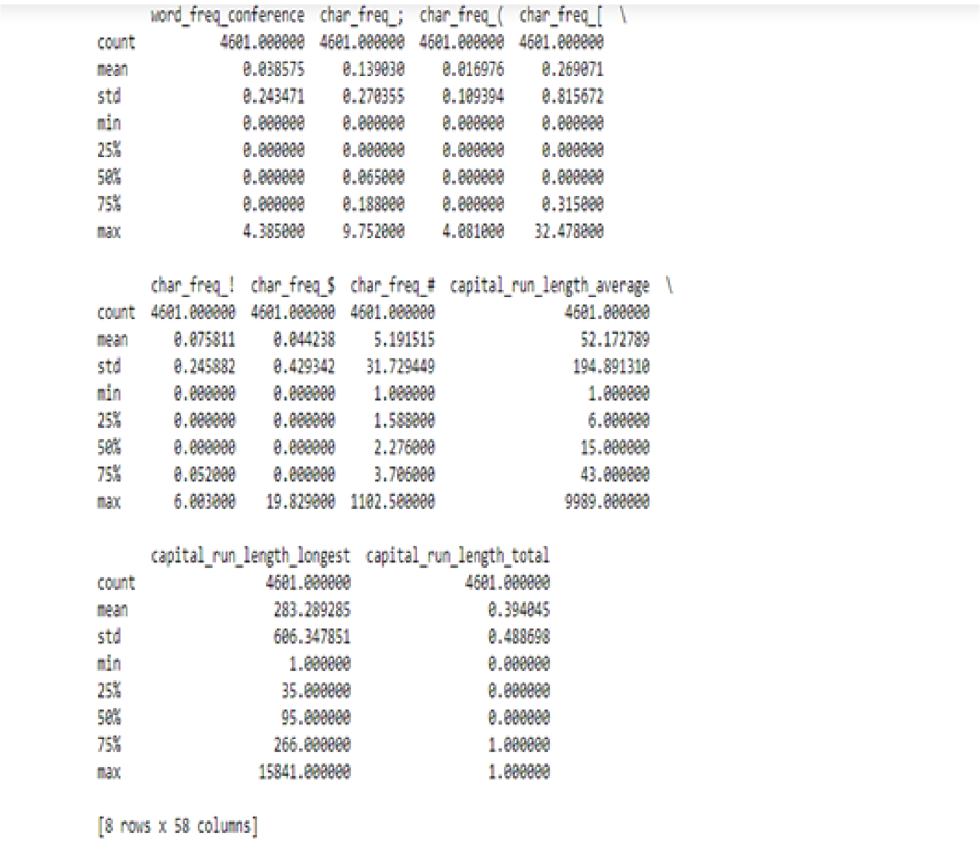
4.1 Logistic regression

Logistic regression is one of the most likely and appropriate algorithm used for classification of datasets. In case of classifying a dataset named as spam base the logistic regression is the most versatile decision based approach for detecting spam emails out of a dataset. Logistic regression performs some basic test on the given distribution of data which involves finding and calculating some statistical domains like mean and standard deviation. It also produces results of the operations like word and character count, max and min operations. After producing and provisioning the statistical and count tests the logistic regression algorithm fetches the outcome of these tests and tends to inter-relate the outcomes.

4.2 Steps followed by the algorithm to classify the spam and non-spam emails

- Selecting and choosing a dataset.

On selecting a data set from MCI machine learning database is an open source website. The dataset contains an email inbox with number of receipts. Some of the receipts have sent the emails repeatedly in a course of time. The fact is to classify the email receipts occurring multiple number of times and to determine the spam emails out of the group.
Printing the contents produced by the logistic regression operations.



```
       word_freq_conference  char_freq_;  char_freq_(  char_freq_[  \
count           4601.000000  4601.000000  4601.000000  4601.000000
mean               0.038575     0.139030     0.016976     0.269071
std                0.243471     0.270355     0.109394     0.815672
min                0.000000     0.000000     0.000000     0.000000
25%                0.000000     0.000000     0.000000     0.000000
50%                0.000000     0.065000     0.000000     0.000000
75%                0.000000     0.188000     0.000000     0.315000
max                4.385000     9.752000     4.081000    32.478000

       char_freq_!  char_freq_$  char_freq_#  capital_run_length_average  \
count  4601.000000  4601.000000  4601.000000                 4601.000000
mean      0.075811     0.044238     5.191515                   52.172789
std       0.245882     0.429342    31.729449                  194.891310
min       0.000000     0.000000     1.000000                    1.000000
25%       0.000000     0.000000     1.588000                    6.000000
50%       0.000000     0.000000     2.276000                   15.000000
75%       0.052000     0.000000     3.706000                   43.000000
max       6.003000    19.829000  1102.500000                 9989.000000

       capital_run_length_longest  capital_run_length_total
count                 4601.000000               4601.000000
mean                   283.289285                  0.394045
std                    606.347851                  0.488698
min                      1.000000                  0.000000
25%                     35.000000                  0.000000
50%                     95.000000                  0.000000
75%                    266.000000                  1.000000
max                  15841.000000                  1.000000

[8 rows x 58 columns]
```

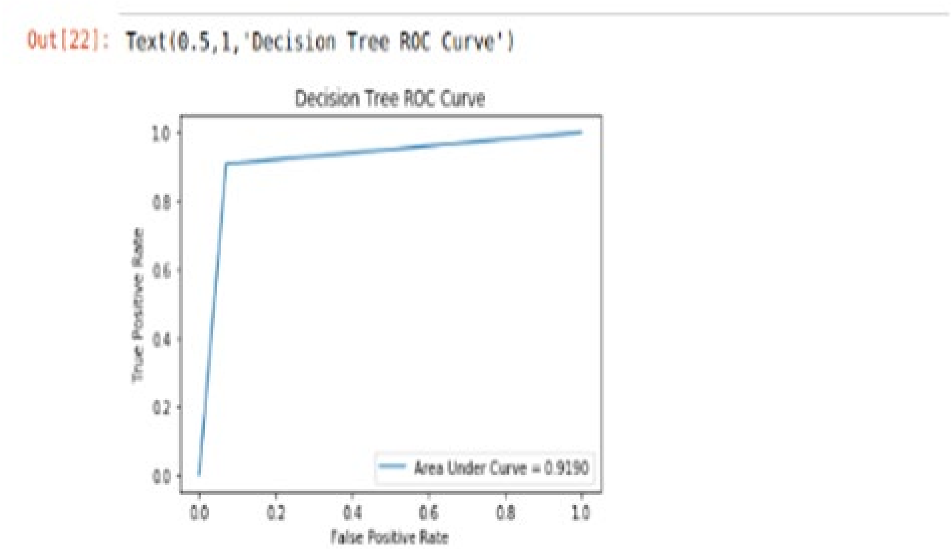
Figure 4.2. View of descriptive statistics 3 with rows and columns.

- Splitting the dataset into test and training dataset.
- Using decision tree to classify emails as spam and non-spam emails.
- Fitting algorithm to training sets.
- Finding the most important predictions.
- Prints the top ten predictors.
- Creating a .dot file to visualize the decision tree. In this step the algorithm will create the decision tree to run tests on the bases of the occurrence of the words.
- Opening spam .dot as a function.
- Accuracy of the model on the test set.
- Prediction on the test set.
- Creation of table for true, false, true positive and true negative, false positive and false negative.

Table 4.1. True false table depicts the role of true/false statements.

Capital_run_length_total	0	1
Row 0		
0	633	48
1	43	427

- Testing of the error rate in the model.
- Finding the area under curve.
- Setting the axes of ROC curve and definition by the ROC curve.



```
Out[22]: Text(0.5,1,'Decision Tree ROC Curve')
```

Figure 4.3. Plotting decision tree ROC curve

4.3 IP address extraction

After finding the groups or sets of spam and non-spam emails, most of the mail servers like Gmail and yahoo provide extra information of the email receipts. After selecting more information the IP addresses of the email receipts could be seen in a dialog box or new window. This sort of information like email type with respective to the IP address of the emails can be handy and useful for tracking the spam-bots and other spam email sending sources. The fact of tracking the spam email sources is one of the basic ideas of the project. Its analogy is simple if a spam-bot or a spam source uses an IP address for sending an spam email, it could be reported or blocked. A mail server provides privileges to the clients for reporting and blocking the emails. One of the key factor in this case is that a list of blocked and reported emails could be used by the firewalls or protocol suits to blacklist such spam sources following the idea of backtracking. The general security means provided by the project is blocking the previous IP address from where an email gets the spam emails. The idea is to making it difficult for the spam sources for sending more and more spam emails from the previous locations. The spam-bots change their location continuously but it could be made difficult by blocking the previous location so that the spam-bot should not use the previous location again in the future.

5. Future scope and conclusion

It has been shown that the spam email sources can be eliminated by the decision based support system. An email user can get exposed to the criminals trying to breach security of an email client. The spams emails are generated for the phishing, advertisements and promotions. The email user can overcome from receiving the spam emails in a bulk amount by following the rules determined by the application designed for that purpose. Anyone can distinguish the spam email receipt and could report or block that particular receipt. Another alternative is to track the receipt by means of the IP addresses and a smart system could instruct the firewall to stop the spam source with the notified IP address, likely a harmful one. Another condition is to update the protocol suite about receiving the emails from an IP address and suggest to block-list that address. As far as organisational emails are considered, it is very difficult for the administrator to remember the name of the recieved email or the keyword of the email by which a prticular emil can be searched. The administratoin is also restricted to block or report the freuently occuring email. An email classifier can be very prominent for such type of classification.

References

[1] Kamoru, B. A., Jaafar, A. B., Murad, M. A. A., Ernest, E. O., & Jabar, M. B. A. Spam Detection approaches and strategies: A phenomenon.

[2] Gao, Q. (2012). Using IP addresses as assisting tools to identify collusions. *International Journal of Business, Humanities and Technology*, *2*(1), 70-75.

[3] Zhong, X. (2014, July). Deobfuscation based on edit distance algorithm for spam filitering. In *2014 International Conference on Machine Learning and Cybernetics* (Vol. 1, pp. 109-114). IEEE..

[4] Khawandi, S., Abdallah, F., & Ismail, A. A Survey ON IMAGE SPAM DETECTION TECHNIQUES. *Computer Science & Information Technology*, 13. *Computer Science & Information Technology*, p.13.

[5] Idris, I., Selamat, A., Nguyen, N. T., Omatu, S., Krejcar, O., Kuca, K., & Penhaker, M. (2015). A combined negative selection algorithm–particle

swarm optimization for an email spam detection system. *Engineering Applications of Artificial Intelligence*, *39*, 33-44.

[6] Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text clustering. In *2005 International Conference on Cyberworlds (CW'05)* (pp. 4-pp). IEEE.

[7] Kamoru, B. A., Jaafar, A. B., Murad, M. A. A., Ernest, E. O., & Jabar, M. B. A. Spam Detection approaches and strategies: A phenomenon.

[8] Puri, S., Gosain, D., Ahuja, M., Kathuria, I., & Jatana, N. (2013). Comparison and analysis of spam detection algorithms. *International Journal of Application or Innovation in Engineering and Management*, *2*(4).

[9] Malarvizhi, R. (2013). Content-based spam filtering and detection algorithms-an efficient analysis & comparison 1.

[10] Krasser, S., Tang, Y., Gould, J., Alperovitch, D., & Judge, P. (2007, June). Identifying image spam based on header and file properties using C4. 5 decision trees and support vector machine learning. In *2007 IEEE SMC Information Assurance and Security Workshop* (pp. 255-261). IEEE.

[11] Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp. 14-16).

[12] Kumari, K. V., & Kavitha, C. R. (2019). Spam Detection Using Machine Learning in R. In *International Conference on Computer Networks and Communication Technologies* (pp. 55-64). Springer, Singapore.

[13] Taooka, Y., Takezawa, G., Ohe, M., Sutani, A., & Isobe, T. (2014). Multiple logistic regression analysis of risk factors in elderly pneumonia patients: QTc interval prolongation as a prognostic factor. *Multidisciplinary respiratory medicine*, *9*(1), 59.

[14] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. *IEEE Access*, *6*, 35365-35381.

[15] S. Journal, R. Statistical, S. Series, and C. A. Statistics, “Algorithm AS 136?: A K-Means Clustering Algorithm Author (s): J . A . Hartigan and M . A . Wong Published by?: Wiley for the Royal Statistical Society Stable URL?: https://www.jstor.org/stable/2346830,” vol. 28, no. 1, pp. 100–108, 2019.

[16] A. Algorithmic, “Machine Learning.”

[17] Z. Burgansky-eliash et al., “Classifiers for Glaucoma Detection?: A Preliminary Study,” pp. 4147–4152, 2005.

[18] D. Mining, K. Discovery, B. Laboratories, and L. Technologies, “A Tutorial on Support Vector Machines for Pattern Recognition,” vol. 167, pp. 121–167, 1998.

[19] C. Jin, L. De-lin, and M. Fen-xiang, “An improved ID3 decision tree algorithm,” no. 1, pp. 1–4.

[20] L. Saitta, “Support-Vector Networks,” vol. 297, pp. 273–297, 1995.

[21] B. Y. D. G. T. Denison, B. A. N. K. Mallick, and A. F. M. Smith, “A Bayesian CART algorithm,” pp. 363–377, 1998.

[22] A. Dey and A. S. Learning, “Machine Learning Algorithms?: A Review,” vol. 7, no. 3, pp. 1174–1179, 2016.

[23] M. Umanol, H. Okamoto, I. Hatono, and H. Tamura, “Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis Systems,” no. 1, 1896.

[24] T. Joachims, “Text Categorization with Support Vector Machines?: Learning with Many Relevant Features,” pp. 2–7.

[25] C. Hsu and C. Lin, “A Comparison of Methods for Multiclass Support Vector Machines,” vol. 13, no. 2, pp. 415–425, 2002.

[26] S. Nasser, R. Alkhaldi, and G. Vert, “A Modified Fuzzy K-means Clustering using Expectation Maximization,” pp. 231–235, 2006.

[27] J. R. Quinlan, DECISION TREES AS PROBABILISTIC CLASSIFIERS. Morgan Kaufmann Publishers, Inc., 2007.

[28] P. Taylor and J. Shao, “Linear Model Selection by Cross-Validation,” no. February 2014, pp. 37–41, 2012.

[29] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” 1999.

[30] P. E. Utgoff, “Incremental Induction of Decision Trees,” vol. 186, pp. 161–186, 1989.

[31] F. Sets, “Induction of fuzzy decision trees,” vol. 69, pp. 125–139, 1995.

[32] Towardsdatascience.com/decision-tree-algorithm.

[33] Ssearchsecurity.techtarget.com

[34] En.wikipedia.org

[35] Talentexplo.com

[36] Docplayer.net

[37] Baadalsg.inflibnet.ac.in

[38] Submitted to Heriot-Watt University

[39] Dspace.vsb.cz

[40] Submitted to UT, Dallas

[41] "International Conference on Computer Networks and Communication Technologies", Springer Nature America, Inc, 2019

[42] whatis.techtarget.com