International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Phishing Detection using Machine Learning Techniques

Meenu , Sunila godara

Abstract: ***Phishing is a type of cyber-crime where spammed messages and false sites allure exploited people to give delicate data to the phishers. The obtained touchy data is along these lines used to take characters or access cash. To battle against spamming, a cloud-based framework Microsoft azure and uses prescient investigation with machine making sense of how to manufacture confidence in personalities. The goal of this paper is to construct a spam channel utilizing various machine learning techniques. Classification is a machine learning strategy uses that can be viably used to recognize spam, builds and tests models, utilizing diverse blends of settings, and compares various machine learning technique, and measure the exactness of a prepared model and figures a lot of assessment measurements. The present study compares the predictive accuracy, f1 score, precession and recall of several machine learning methods including Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), and Neural Networks (NNet) for predicting phishing emails and improves logistic regression technique by using feature selection methods and improves the accuracy to detect phishing.***

Keywords: DT , LR , NN, Phishing, SVM.

I. INTRODUCTION

Phishing technique is used to steel personal information using fake email messages and for the purpose of identifying the theft, this is done by sending emails to gain access to personal confidential information. Phishing email crime is increasing very fast for stealing personal information. The person who response the receiving email, or enter the personal information into email then the data of the person is at risk.

a) Machine learning phase:

Microsoft Azure platform provides tools for machine learning. In these experiments, the two class boosted decision tree and the two class support vector machine (SVM) were used as spam classifiers. The decision tree is

Mainly used in data mining . It has the ability to create a model that foreshows the value of a target variable based on various input variables. The SVM is a supervised learning model that has learning algorithms and the ability to analyze data for classification. Given a set of training examples, SVM can decide whether an email belongs to the "spam" or "good" email category.

Separate datasets were generated to train and test the models. First, the data was split into training and test data. Then, the models were trained and evaluated. By using the Azure machine learning studio, we were able to try decision tree and SVM and compare our results. This type of experimentation assisted in finding the best solution to the study problem. The test data that resulted was used to score the trained models.

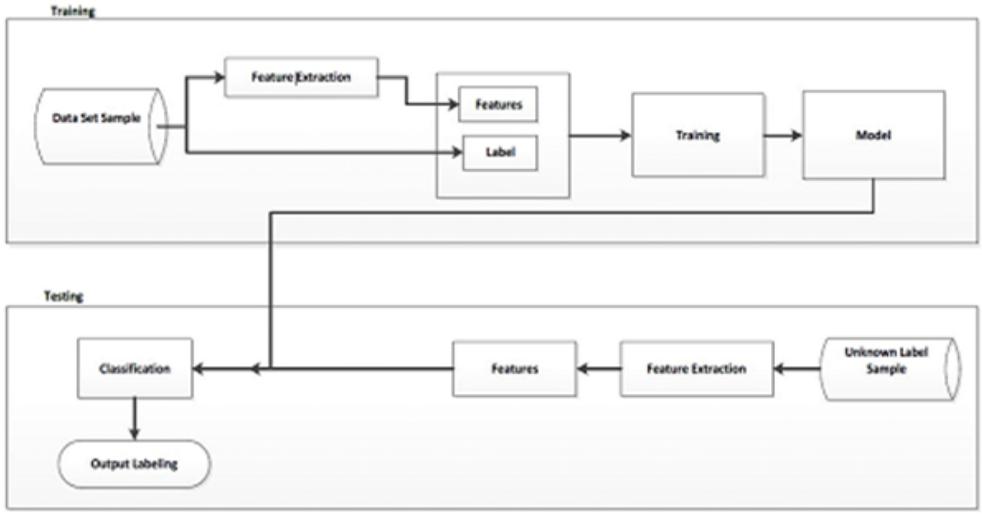

Fig: 1: Automated Email Phishing Detection

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Meenu , computer science , guru jambeshwar University of science and technology , hisar , Haryana , India . Email: meenubhobhia@gmail.com.
Sunila godara , computer science department , guru jambeshwar University of science and technology ,hisar ,Haryana ,India.

These techniques extract values from email from various predefined features to classify the email either spam or ham.

b) Phishing:

Phishing is an illicit endeavor that adventures both social building and specialized misdirection to obtain touchy secret information (e.g. government managed savings number, email address, passwords, and so on.) and money related record certifications. Phishing includes spam messages camouflaged as authentic with a subject or message intended to trap the casualties into uncovering classified data. In misleading phishing, email warnings from charge card organizations, security offices, banks, suppliers, online installment processors or IT overseers are used to abuse the clueless open. The notice urges the beneficiary to direly enter/refresh their own information [2].

Phishing Emails Detection Techniques

Filters are developed to prevent phishing and manage traditional techniques like authentication protection and some other modern various data mining techniques

1. Traditional Method for phishing detection

This method falls in two categories, one is authentication protection and other is network level protection. Network level protection includes two type of filters, white -list filter and black-list filter which are used to prevent phishing by blocking IP address and domain from network. There are also a rule based filter and pattern matching filter.

- ***Black list Filter***

This filter provide protection at network layer by classifying receiver email DNS address ,IP address or sender address .from the email header detail are extracted and compare it with predefined list , if the data is match then the email is rejected otherwise accepted .This technique provide security.

- ***White list Filter***

This filter also provides network level protection. This filtering technique compare the email data with predefined list containing IP address and static IP addresses of legitimate domain . In this technique only those emails are allowed to access the user inbox from the network which is match with the list.

- ***Pattern Matching filter***

This filter provide protection at network level and include some specified filters like string ,word , text and various character set that are available in the email content . This filter is used to classify the email either spam or ham by searching the email in pattern list and if the received email include large amount of banned text or words then this will give invaluable result.

To provide security to user and domain level authentication protection is used and this protection is created for domain level which works for email server. For user level protection before sending message user must provide authentication.

- ***Email verification***

Email confirmation is a client level verification system. This system requires confirmation from sender and collector. At the point when the sender get the message, email will guarantee and named ham and afterward went to the recipient inbox generally delegated ham and keep it from the getting to in the inbox. This sifting gives great precision to identify spam in the email but it will take a lot of time because before receiving a message receiver response is necessary. In this process if verification process generate traffic on the network then the email is at risk or it may be loss.

- ***Password Filter***

Password filter provide user level authentication protection .it allow to receive any email in the email address ,subject, header field , if the filter detect a wrong password and not detect the password then email is rejected .for creating the password the user of this filter communicate with each other to set the password ,the password was not created by default . This filter has some limitation when some emails are lost when the password is not recognized or the process will take more time to detect password

2. Automated Methods

Automatic classifier on machine learning is used to classify the received mail is spam or ham.

- ***Logistic regression***

This is a type of automated method and applies a linear model to predict binary data i.e. 0 or 1. This method is easily interpreted and gives good result by classifying email as spam or ham. This is a simple method for classifying email as spam or ham.

Classification and Regression Trees (CART)

This method is utilized to speak to the tree which is parts utilizing two hubs .this will make a twofold tree and is utilized for complex connection between the factors as opposed to direct connection.

A tree is made to parts the indicator space into various gatherings and this parts is rely upon apportioning principles related to interior hubs of the hubs ,each gathering is related to inside hub of the tree. This model will produce a parallel tree for the perplexing connection and simple read collaboration among indicators is given. It additionally make hard to foresee the added substance impact because of its tremendous.

- ***Decision Trees Filter (DT)***

This channel is a graphical model for order. Decision tree contain nodes and arrows and it is initialize from the root node. If-then rules are within each node in the network. Arrow represents which node referred to next. Tree will also contain various classifier stage and internal nodes.

Various terminologies are used in decision tree are:

- Root hub: base hub is called root hub from which tree is instate.
- If-then principle: every hub inside the system contains If-then guideline, a class and an element.
- Arrow: next edge is eluded by utilizing bolt.
- Leaf hub: tree closes with leaf hub or the eliminator.
- To produce a tree different calculations are incorporated ID3 model to ascertain entropy data to assess the objective worth .in C4.5 calculation tree will create sub trees in which every hub of the tree has a parent hub and furthermore prompts a kid hub. Furthermore, the tree closes with ending hub that speaks to the objective yield of the issue.

Support Vector Machine (SVM)

This technique is used in medical for diagnosis of diseases, text recognition, for classification of image and in the other fields. This will partition the data into two categories using fixed rule, quadratic equation and statistic.

Separating hyper plane is used for the binary classification of the data and minimizes the space of the margin on the basis of kernel function. This technique is used to find the best solution of the problem. This technique is fails in analyzing the big data.

Various feature selection methods are:

The Filter Based Feature Selection module provides multiple feature selection algorithms to choose from, such as Pearson's or Kendall's correlation, mutual information, fisher scores, and chi-squared values. In this we use chi square method for feature selection

- **Pearson Correlation:** Label can be text or numeric. Features must be numeric
- **Mutual Information:** Labels and features can be text or numeric. Use this method for computing feature importance for two categorical columns.
- **Kendall Correlation:** Label can be text or numeric but features must be numeric
- **Spearman Correlation:** Label can be text or numeric but features must be numeric
- **Chi Squared:** Labels and features can be text or numeric. Use this method for computing feature importance for two categorical columns.
- **Fisher Score:** Label can be text or numeric but features must be numeric.

II. STUDY AND REVIEW OF LITERATURE

MEENA, P et al. [1] proposed system that helps to detect and prevent the phishing mails using two techniques named as Phishzoo and MLAPT (Machine Learning Anti-Phishing System). Phishzoo it is a technique used to detect the phishing websites by the way of their appearances and MLAPT technique is for preventing the phishing mails on the system. With the help of these two techniques the user can maintain their personal details on the social networking system.

Henry [4] presents some of the techniques that phishers use to attack the user in brief. Also, there are many techniques developed to detect phishing and protect user's data from being stolen. So this paper mainly discusses schemes to detect Phishing attack.

Jacobsonet al. [5] introduces tools to model and describe phishing attacks, allowing a visualization and quantification of the threat on a given complex system of web services. They utilize new model to portray some new phishing assaults, some of which have a place with another class of mishandle presented thus: the setting mindful phishing assaults. They depict methods for utilizing the model we acquaint with evaluates the dangers of an assault by methods for monetary investigation, and strategies for protecting against the assaults portrayed. Watchwords: setting mindful, character

Chhikara et al. [6] presents brief data about phishing, its assaults, steps that clients can take to defend their secret data. This paper additionally demonstrates an overview led by net craft on phishing.

Abu-Nimeh et al. [7] show consider takes a gander at the farsighted precision of a couple of AI systems including Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNet) for anticipating phishing messages. An enlightening list of 2889 phishing and true blue messages is used as a piece of the comparable assessment. In addition, 43 features are used to plan and test the classifiers.

Kumar et al. [8] utilized TANAGRA information mining apparatus on a tested spam dataset to assess the productivity of the messages classifier where a few calculations were applied on that informational index. Toward the end, the highlights determinations by Fisher spam channels and separating accomplished better characterizations. After fisher sifting has accomplished over 99% precision in recognizing spam, the tree characterization calculation was applied on pertinent highlights.

Ping et al. [47] proposed a b-bit hashing mind direct learning calculation, for example, straight SVM or strategic relapse to explain huge scale and high dimensional measurable learning task .they contrast b-bit hashing and the Count-Min (CM) and Vowpal Wabbit (VW) calculations, which have basically indistinguishable changes from arbitrary projections .and their correlation delineate that hashing is more exact than VW for the parallel information.

Azad et al. [42] concentrated on testing distinctive existing calculations as far as their precision, for example, Naive Bayes, strategic relapse, and bolster vector machine (SVM) classifiers. He utilized sack of words and increased pack of words models. By and large, the tried classifiers accomplished high outcomes showing an exactness pace of 95% with the SVM with the straight portion and Bayes beating different classifiers, as they just missed 10 and 2.66 percent of phishing messages individually.

III. EXISTING TECHNIQUE OF MACHINE LEARNING TO DETECT PHISHING

Human capacity is limited and he/she cannot prevent and detect all the phishing but the machine is intelligent and this can do all this work fast and prevent from intrusion. Therefore machine learning is the best technique to solve the problem. Phishing is detected by using various machine learning techniques.

a) MACHINE LEARNING

This is a field of artificial intelligence and it has ability to learn without explicitly programmed. Various machine learning techniques are supervised learning, unsupervised learning and reinforcement learning.

Machine learning types

Types of machine learning techniques are:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

1) Supervised learning

This is a machine learning type which is similar to teacher from which human learns. Teacher gives good example to the students and the student derive rule from this example.

Types of supervised learning

- ***Regression:*** predict the *continuous-response* value.
- ***Classification:*** predict the *categorical response* value where the data is separated into "**classes**"

2) Unsupervised learning-

In this learning algorithm learns from plain example without any associated response leave on algorithm to determine pattern. This algorithm restructures the data into other forms such as new features which represent a class and new series of unrelated data. It is useful to providing new useful input to the algorithm. The training data does not include Targets here so we don't tell the system where to go; the system has to understand itself from the data we give.

Types of unsupervised learning are:

- ***Clustering:*** *This is a type of problem where we* group similar things together.

3) Reinforcement learning

This type of machine learning occurs when the algorithm does not contain any label like unsupervised learning. Reinforcement learning is connected to the application for which algorithm must take decision. This algorithm learns by trial and error method.

b) Classification techniques:

Characterization systems can be utilized to foresee results i.e. spam or ham .different methods that are utilized to order spam or ham are two class calculated relapse procedure, two class helped choice tree, two classes bolster vector machine, and two class neural systems. Order is a machine learning technique that is utilized to decide the sort, or class of a thing.

For instance, you can utilize grouping to

- Classify email as spam or ham.
- Determine whether a patient's test report is positive or negative.

Two class logistic regression

This technique is used to create phishing detection model which predict only two outcomes that is spam or ham. This Is a statistical or supervised learning method and for this classification technique to train a model we provide dataset .this technique is used to predict the probability of the result .this technique use logistic function to predict the probability by fitting the data set .this technique is used for two class problems that contain two values and a data set containing label is used to train the model.

Two class Boosted Decision tree

Two-Class Boosted Decision Tree makes an AI model that relies upon the helped decision trees figuring. A helped decision tree is a troupe learning system in which the subsequent tree changes for the mix-ups of the principle tree, the third tree amends for the goofs of the first and second trees, and so forth. Desires rely upon the entire social affair of trees together that makes the figure. Supported decision trees are the easiest methods with which to get top execution on a wide variety of AI endeavors. The model by then picks the perfect tree using an emotional differentiable hardship work.

- Decision hub: this hub shows choice to be made.
- Leaf hub: shows ultimate result of the choice way for example spam or ham
- Branch: each branch demonstrates conceivable result.

Each leaf of the tree is set apart with a class and an upheld decision tree is an outfit learning procedure in which the subsequent tree overhauls for the slip-ups of the principle tree, the third tree corrects for the bungles of the first and second trees, and whatnot. Estimates rely upon the entire social occasion of trees together that makes the desire.

Two class Support vector machine

Support vector machines (SVMs) are a particularly asked about class of directed learning systems. This particular execution is fit to the desire for two possible outcomes, in perspective on either relentless or unmitigated elements.

This is an especially investigated class of managed learning techniques. This particular execution is fit to estimate of two possible outcomes, considering either relentless or absolute variables. In the wake of describing the model parameters, set up the model by using one of the readiness modules, and giving

a named dataset that consolidates an imprint or result area. SVM models have been used in various applications, from information recuperation to substance and picture gathering.

Two class neural system

A neural framework is a course of action of interconnected layers. The data sources are the essential layer, and are related with a yield layer by a non-cyclic outline included weighted edge. Most insightful tasks can be refined viably with only a solitary or a few covered layers. Regardless, continuous investigation has shown that significant neural frameworks with various layers can be very convincing in complex assignments, for instance, picture or talk affirmation. The dynamic layers are used to show growing degrees of semantic significance. The association among wellsprings of data and yields is discovered from setting up the neural framework on the data. The heading of the chart proceeds with the commitments through the covered layer and all centers in a layer are related by the weighted edges to. A neural framework is a course of action of interconnected layers.

A neural system is a lot of interconnected hubs .The principal layer is information layer which is associated with the concealed layer and this shrouded layer is associated with the yield layer.

- Input layer: this layer speaks to the information.
- Hidden layer: this layer speaks to the transitional computation advertisement figure limit held up total of the information.
- Output layer: speak to the yield.

Comparison of existing technique

Classification technique	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.941	0.9567	0.9365	0.938
Neural Network	0.9431	0.9601	0.9430	0.944
Decision Tree	0.939	0.9557	0.933	0.936
Support Vector Machine	0.886	0.8885	0.904	0.931

Table 1: comparison of various machines learning Technique

Following figure compare the various machine learning techniques.

IV. EVALUATION APPROACH:

This section describes about the data set and also describes evaluation metrics that are used in comparison.

Data set description:

Data set contains 2,000 marked messages for preparing and 100 named messages for testing. Each message is marked either spam or ham (not spam).

Evaluation metrics:

By utilizing accuracy, precession, recall and F1 score metrics look at different models and discover which show accomplishes the best outcome for an order of spam or ham.

- **Accuracy**: this will gauge the level of the right consequence of an order to demonstrate.
- **Precision:** this is a level of genuine forecast that is right.
- **Recall:** this s a small amount of positive occurrence that was anticipated as positive and gives the entire right outcome returned by demonstrating.
- **F-Score:** it is figured as the heaviness of ac

V. PROPOSED HYBRID APPROACH TO DETECT PHISHING USING MACHINE LEARNING TECHNIQUE

Machine learning techniques are used to detect phishing. Proposed approach to detect phishing using machine learning techniques is:

1. Preprocess the information to evacuate unneeded data
2. Select segments in an informational collection to choose section which we need and afterward alter metadata
3. Now change a flood of English content into a lot of highlights spoke to as numbers.
4. Then pass this hashed list of capabilities to an AI calculation to prepare a book investigation model.
5. Feature determination the way toward applying measurable tests to inputs, given a predetermined yield, to figure out which segments are increasingly prescient of the yield.
6. Split informational index to partition a dataset into two unmistakable sets. Separate information into preparing information and testing information.
7. Two class Logistic relapse system exactness is improved by changing a few parameters
8. Tune Model Hyper parameters are utilized to decide the ideal hyper parameters for a given AI model. Manufactures and tests numerous models, utilizing various mixes of settings, and thinks about measurements over all models to get the blend of settings.
9. Score informational index to create a lot of measurements utilized for assessing the model's precision.
10. Evaluate Model to gauge the precision of a prepared model and registers a lot of assessment measurements.

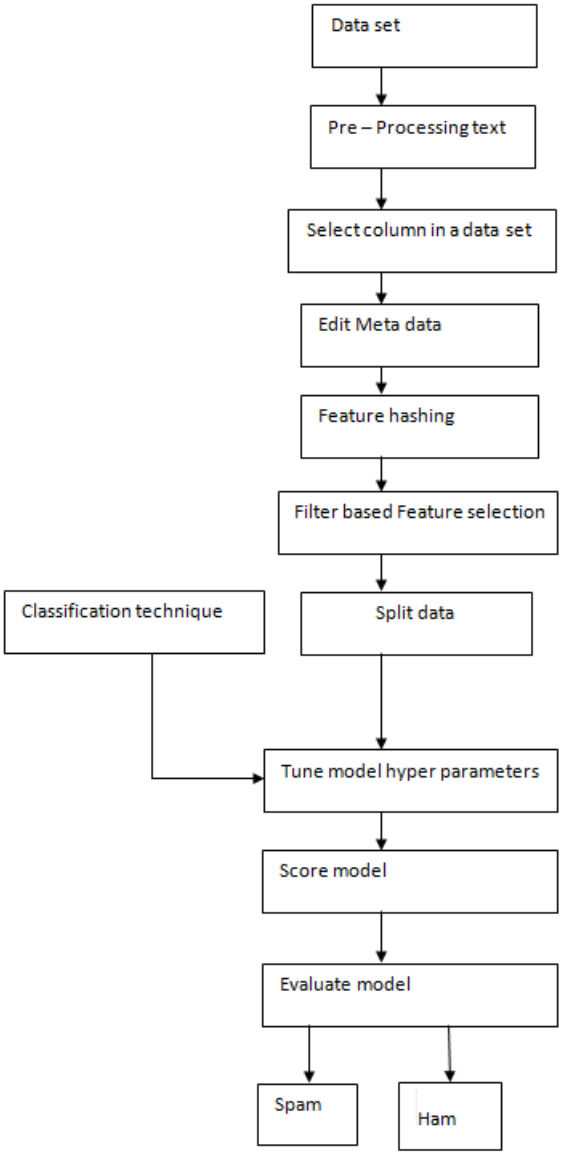

Fig 2: Proposed approach

VI. EXPERIMENTAL RESULTS:

This section demonstrate experimental studies to investigate the predictive accuracy,f1 score, precession and recall of NN, LR, DTand SVM by using various feature selection methods and improve the accuracy of logistic regression technique by using fisher score feature selection method .compare existing machine learbning technique and improve logistic regression technique.

Classification technique	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.941	0.9567	0.9365	0.938
Neural Network	0.9431	0.9601	0.9430	0.944
Decision Tree	0.939	0.9557	0.933	0.936
Support Vector Machine	0.901	0.91	0.89	0.896
Improved Logistic Regression	0.955	0.9608	0.9489	0.95

Table2: Comparison of various machine learning techniques

Result of two logistic regression techniques is improved by using feature selection method. Fisher score method is used to select the feature. This method for feature selection is simple, feasible and time saving. This is an efficient procedure and maximizes the likelihood by getting successively closer and closer to the maximum by taking another step. Selecting features by Fisher score can improve the accuracy of subsequent procedure.

Accuracy of logistic regression is 0.941, F1 score is 0.9567, Precession is 0.9365 and recall is 0.938.

Accuracy of neural network is 0.9431, F1 score is 0.9601, Precession is 0.9430 and recall is 0.944.

Accuracy of decision tree is 0.939, F1 score is 0.9557, Precession is 0.933 and recall is 0.936.

Accuracy of support vector machine is 0.901, F1 score is 0.91, Precession is 0.89 and recall is 0.896.

Accuracy of logistic regression is 0.955, F1 score is 0.9608, Precession is 0.9489 and recall is 0.95.

Table 1 shows the comparison between accuracy, F1 score, Precision and recall of logistic regression, neural network, decision tree, and support vector machine and improve the logistic regression technique by using fisher score feature selection method.

Fisher scores

the Fisher scores of all the features, and for a given threshold θ, feature fi is selected if F(fi) > θ, otherwise, if F(fi) ≤ θ, feature fi will not be selected. Selecting features by Fisher score can improve the accuracy of subsequent procedure (such as classification and prediction), and the process is simple, feasible and time saving.

Advantages of fisher score method:

- Selecting features by Fisher score can improve the accuracy of subsequent procedure
- the process is simple,
- feasible
- Time saving.
- it maximizes the likelihood by getting successively closer and closer to the maximum by taking another step (an iteration)
- It is an efficient procedure.

Compare various machine learning techniques like logistic regression , neural network , decision tree and support vector machine by using various feature selection methods like Pearson correlation, chi squared method and Kendall correlation .

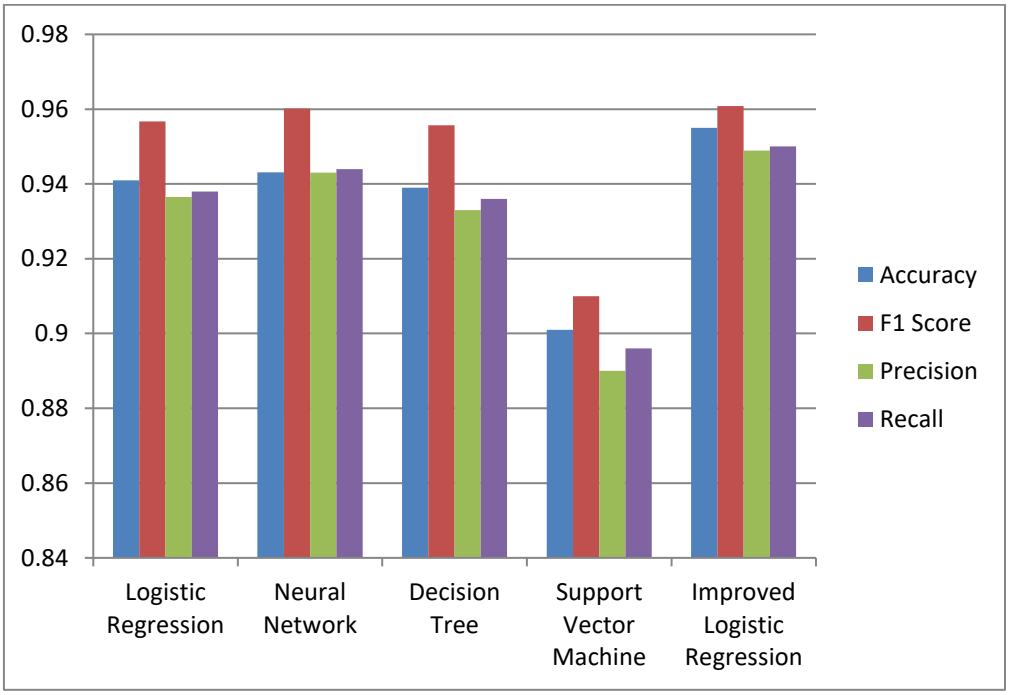

Fig 3: Comparison of accuracy, f1score, precision and recall of various machine learning techniques

Following figure compare the accuracy, f1 score, precession and recall of various machine learning Techniques accuracy of improved logistic regression technique is 0.955, f1 score is 0.9608, precession is 0.9489 and recall is 0.95.

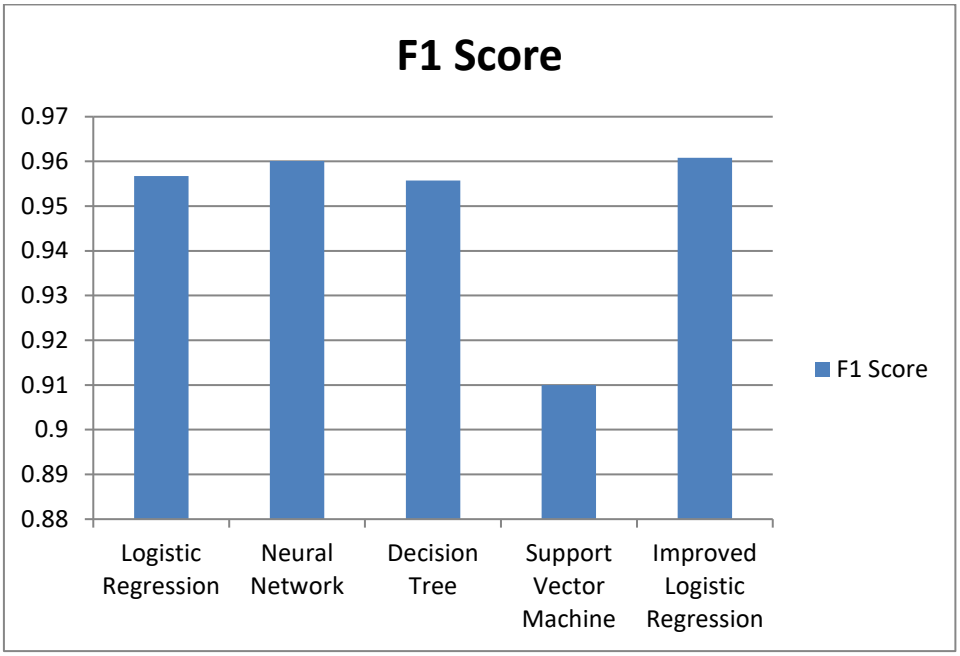

Fig 4: comparison of F1 score of various machine learning technique

Fig compares the F1 score of various machine learning techniques. F1 score of logistic regression is 0.9567, Neural network is 0.9601, decision tree is 0.9557, support vector machine is 0.91 and improved logistic regression is 0.9608.

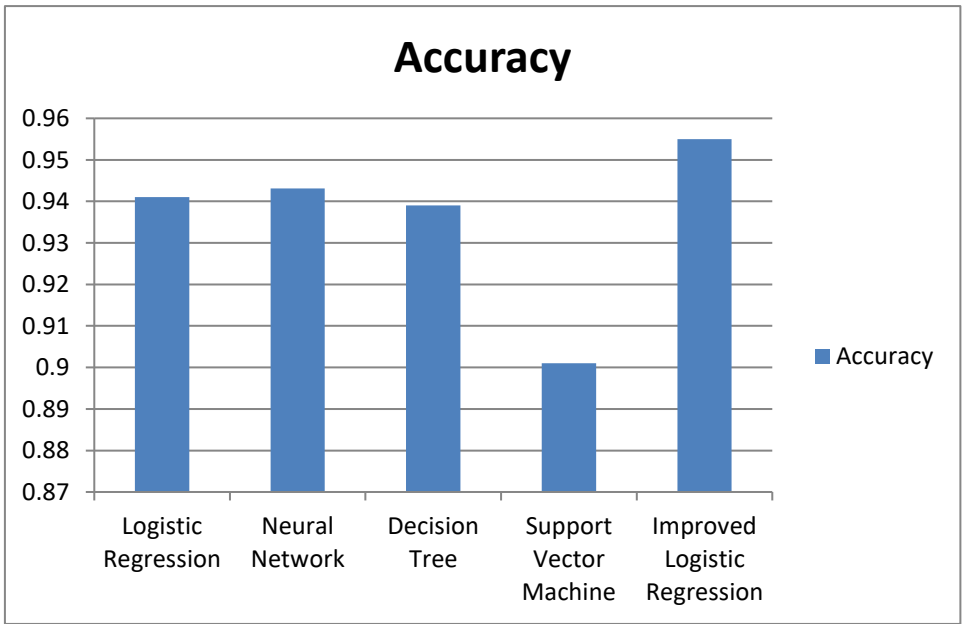

Fig 5: comparison of accuracy of various machine learning techniques

Fig compares the accuracy of various machine learning techniques. accuracy of logistic regression is 0.941, neural network is 0.9431, decision tree is 0.939, support vector machine is 0.901 and improved logistic regression is 0.9555.

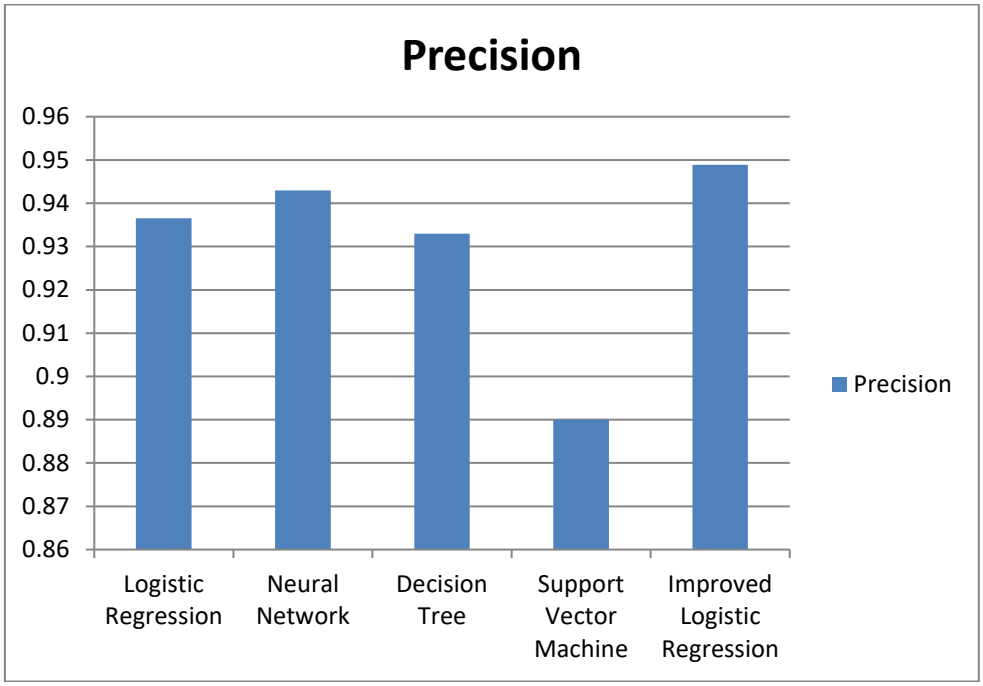

Fig 6: comparison of precession of various machine learning techniques

Fig compare the Precession of various machine learning technique precession of logistic regression is 0.9365, neural network is 0.9430, decision tree is 0.933, support vector machine is 0.89 and improved logistic regression is 0.9489.

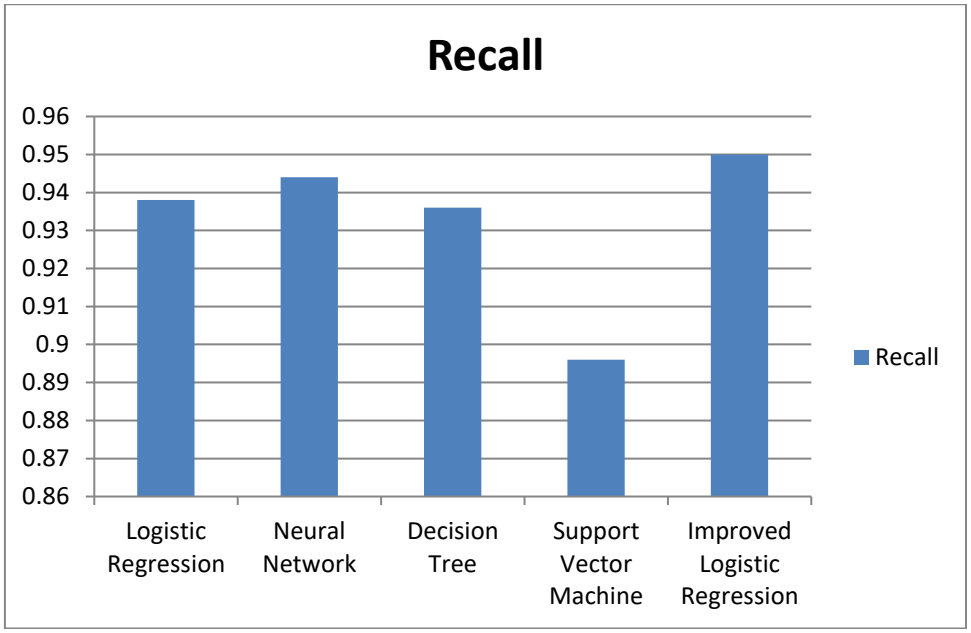

Fig 7: comparison of recall of various machine learning techniques

Fig compare the recall of various machine learning technique recall of logistic regression is 0.938, neural network is 0.944, decision tree is 0.936, support vector machine is 0.896 and improved logistic regression is 0.95.

VII. CONCLUSIONS AND FUTURE SCOPE

Conclusion

This investigation proposes framework that utilization machine learning systems to beat the spam issue. A model of the framework has been produced on the Azure stage and the conduct of email servers has been examined. Develop a phishing detection model by using various data mining techniques to enhance the phishing detection accuracy and a feature selection method are also used to increase the accuracy of the classification model by selecting best feature and find best result. Vow pal Wabbit is a fast machine learning framework used by Feature Hashing ,which is used to hashes feature word into n memory indexes ,by using hash functions Finally, the comparison various machine learning techniques like two class logistic regression technique and two class boosted decision tree (DT) ,two class neural network(NN) and two class support vector machine (SVM) and improved logistic regression is proposed to detect spam .

Future scope

Feature selection technique need more improvement to develop a new phishing detection technique that give best accuracy , So recommend to develop a new tool to extract new features from new raw emails to improve the accuracy. There is need to find out the best parameter which improve the accuracy of various classification technique and give best result in detection of spam or ham

REFERENCES

1. meena, p., m. kavitha, s. jeyanthi, and cpnijithamahalakshmi. "phishing prevention using datamining techniques." *International Journal of Pure and Applied Mathematics* 119, no. 10 117-123, 2018.
2. Meenu , Sunila godara"An enhanced phishing email detection model using machine learning techniques"*international journal of emerging technologies and innovative research 11* ,vol 5,pp523-529 , november 2018.
3. Meenu , Sunila godara"Analysis of various Machine Learning Techniques to Detect Phishing Email: *International Journal of Computer Applications* vol 178(38):4-12 · August 2019 .
4. Henry, Azriel, and JwalantBaria. "Phishing attacks and Schemes to detect Phishing: A *Literature Survey*." 2017.
5. Jakobsson, Markus. "Displaying and counteracting phishing assaults." *In Financial Cryptography*, vol. 5. 2005.
6. Chhikara, Jyoti, RituDahiya, NehaGarg, and Monika Rani. "Phishing and hostile to phishing methods: Case ponder." *International Journal of Advanced Research in Computer Science and Software Engineering 3*, no. 5, 2013.
7. Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "An examination of machine learning systems for phishing recognition." *In Proceedings of the counter phishing working gatherings second yearly eCrime specialists summit*,ACM, pp. 60-69, 2007.
8. Kumar, R. K., Poonkuzhali, G., and Sudhakar, P. Similar investigation on email spam classifier utilizing information mining procedures. *In Proceedings of the International Multi Conference of Engineers and Computer Scientist* Vol. 1, pp. 14-16,march-2012.
9. Li, Ping, Anshumali Shrivastava, Joshua L. Moore, and Arnd C. König. "Hashing algorithms for large-scale learning." *In Advances in neural information processing systems*, pp. 2672-2680. 2011.
10. Azad, B. Recognizing Phishing Attacks.

AUTHOR PROFILE

Meenu, Mtech (computer science) at guru jambeshwar University of science and technology , hisar , Haryana , India . She has published three research papers in international journal and her research interest are in the area of machine learning techniques.

Dr. Sunila Godara, is Ph.d holder and has 14 year teaching experience. She has published various research paper in international journal and conferences. she is Associate professor at guru jambeshwar University of science and technology , hisar , Haryana , India