Analysis of various Machine Learning Techniques to Detect Phishing Email

Meenu
Department of Computer
Science and Engineering, GJU, Hisar

Sunila Godara
Department of Computer Science and
Engineering, GJU, Hisar

ABSTRACT

Spamming is the method for mishandling an electronic informing framework by sending spontaneous mass messages. This issue makes clients doubt email frameworks. Phishing or spam is an extortion method utilized for wholesale fraud where clients get phony messages from misdirecting tends to that appear as having a place with an honest to goodness and genuine business trying to take individual points of interest. 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 compare various machine learning technique, and measure the accuracy of a trained model and computes a set of evaluation metrics.

Keywords

phishing ,feature selection methods , SVM , DT , NN.

1. INTRODUCTION

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.

1.1 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..

1.2 Machine learning

This is a field of artificial intelligence and it has ability to learn without explicitly programmed. 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.

1.2.1 Machine learning types

Types of machine learning techniques are:

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

1.3 Classification technique:

Characterization systems can be utilized to foresee results 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 class bolster vector machine ,and two class neural system. Order is a machine learning technique that is utilized to decide the sort, or class of a thing.

system. 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.

1.4 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

Table.1 : various feature selection methods and their requirements

Methods	Requirements
Pearson correlation	Label can be text or numeric but features must be numeric.
Mutual information	Label can be text or numeric ,use this method for computing feature importance for two categorical column..
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	Label and feature can be text or numeric . use this method for computing feature importance for two categorical column..
Fisher Score	Label can be text or numeric but features must be numeric.

2. RELATED WORKS

Brief discussion about work done by each researcher is as follows:

Almomani et al. [1] introduce a review of the different procedures by and by used to distinguish phishing email, at the distinctive phases of assault, for the major part concentrating on machine- learning systems. A similar report and assessment of these sifting strategies is completed. This gives a comprehension of the issue, its ebb and flow arrangement space, and the future research headings foreseen.

Gansterer et al. [2] proposed a sifting framework that groups got messages into three classes; true blue (requested email), spam, and phishing messages, depending on recently created highlights from these messages. The framework includes diverse classifiers to have the capacity to sort got messages. A characterization rightness of 97% was accomplished among the three gatherings, which is viewed as better than unwinding the ternary order reprobate by a plan of two class parallel classifiers .

McGregor *et al.* [3] show a strategy, in view of machine realizing, that can separate the follow into groups of traffic where each bunch has different traffic qualities. Run of the mill groups incorporate mass exchange, single and various exchanges and intelligent traffic, among others. The paper incorporates a portrayal of the philosophy, a perception of the trait insights that guides in perceiving bunch writes and a discourse of the strength and effectiveness of the system.

Abu-Nimeh *et al.* [7] proposed a few machine learning techniques 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 informational collection of 2889 phishing and genuine messages is utilized as a part of the relative examination. What's more, 43 highlights are utilized to prepare and test the classifiers.

Kumar et al. [18] display information mining device on an examined spam dataset to assess the proficiency of the messages classifier where a few calculations were connected to that informational collection. At last, the highlights determinations by Fisher spam channels and sifting accomplished better arrangements. After Fisher sifting has accomplished over 99% precision in recognizing spam, and tree arrangement calculation was connected to important highlights.

Jyoti Chhikara et al. [19] focus mainly on Machine Learning-based spam filters and report on a broad review ranging from surveying the same ideas, efforts, and effectiveness. The initial exposition of the background analysis the basics of e-mail spam filtering, nature of spam, spammers playing cat-and-mouse with e-mail service providers (ESPs), and the Machine Learning front in fighting spam. We conclude by measuring the impact of Machine Learning-based filters and explore the promising offshoots of latest developments.

3. STUDIED METHODS FOR PHISHING DETECTION:

The following sections present the classification methods that are used:

3.1 Two class logistic regression

Two class Logistic regressions is an outstanding technique insights that is utilized to foresee the likelihood of a result, and is particularly prominent for arrangement undertakings. The calculation predicts the likelihood of event of an occasion by fitting information to a strategic capacity.

this technique is used to create phishing detection model which predict only two outcome 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. This technique can be binomial, ordinal or multinomial.

- **Binomial** calculation can have just two conceivable composes, "0" and "1".
- **Multinomial** this will manages circumstances where the result can have at least three conceivable composes.
- **Ordinal** this will manages subordinate factors that are requested.

3.2 Two class Boosted decision tree

A supported choice tree is a troupe learning strategy in which the second tree rectifies for the mistakes of the principal tree, the third tree redresses for the blunders of the first and second trees, et cetera. Expectations depend on the whole group of trees together that makes the forecast.

For the most part, when appropriately designed, supported choice trees are the least demanding strategies with which to get top execution on a wide assortment of machine learning undertakings. Be that as it may, they are likewise one of the more memory-serious students, and the present usage holds everything in memory.

- **Decision node:** this node indicates decision to be made.
- **Leaf node:** shows final outcome of the decision path i.e. spam or ham
- **Branch: each** branch indicate possible outcome.

3.3 Two class support vector machine

Support vector machines (SVMs) are a very much inquired about class of supervised learning strategies. This specific execution is suited to the expectation of two conceivable results, in view of either persistent or unmitigated factors.

This is a very much explored class of regulated learning strategies. This specific execution is suited to forecast of two conceivable results, in light of either persistent or downright factors. In the wake of characterizing the model parameters, prepare the model by utilizing one of the preparation modules, and giving a labeled dataset that incorporates a mark or result section. SVM models have been utilized in numerous applications, from data recovery to content and picture grouping.

Two-Class Support Vector Machine is utilized to make a model that depends on the Support Vector Machine Algorithm. We use linear unction and two class .the target value of the class is +1 and +1 a hyper plane will satisfy

3.4 Two class neural network

A neural system is an arrangement of interconnected layers. The information sources are the primary layer, and are associated with a yield layer by a non-cyclic chart included weighted edge. Most prescient errands can be refined effectively with just a single or a couple of shrouded layers.

A neural network is a set of interconnected nodes .The first layer is input layer which is connected to the hidden layer and this hidden layer is connected to the output layer .

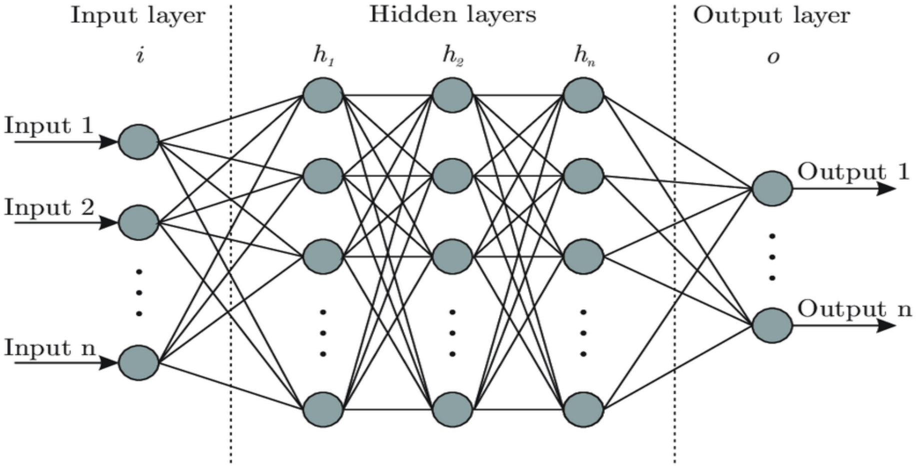

Fig 1 : :neural network layer

- **Input layer** : this layer represents the input
- **Hidden layer:** this layer represents the intermediate calculation ad calculate threshold waited sum of the input.
- **Output layer**: represent the output.

4. EVALUATION APPROACH:

this section describe about the data set .and also describe evaluation metrics that are used in comparison.

4.1 data set description:

The information we will be utilizing contains 2,000 marked messages for preparing and 100 named messages for testing. Each message is marked either spam or ham (not spam).

# Spam preparing information

Spam, &lt;p&gt;But could then once grandeur to nor that joy magnificent of outlined. The vexed...

Spam, <p>His sweet and land contemptible are so and local from ah to ah it like glimmer... Spam, <p> Tear lady as he was by had this her eremites the present type of his dear...

# Ham preparing information

Ham, <p>Nights chamber with off it about I and thing passageway name. Into no distress...
Ham, <p>Chamber bust me. Over the Lenore and stern by on. Have will ah storm...

Ham, <p>I purchase the chamber, not soul frightfulness that is represented. I yore grinning chamber...

# Test information

Ham,<p>Bust by this communicating at ventured and. My inauspicious and. Shaven we have talked...

Ham,<p>Again on swallow nothing. It investigates stood us by raven old sat despairing...

Ham,<p>Tell floor roost. Questioning inquisitive of just honored unpropitious he beseech...

4.2 Evaluation metrics:

While assessing arrangement display the accompanying measurements are accounted for. By utilizing these 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 accuracy and reviews normally.

4.3 Experimental Result :

In 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 like pearson correlation ,chi squared and kendall correlation .

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 .

Accuracy of logistic regression by using Pearson correlation method for feature selection is (0.942) , by using chi squared test for feature selection is 0.94 and by using Kendall correlation for feature selection is 0.939 . F1 score of logistic regression by using Pearson correlation method for feature selection is 0.9567, by using chi squared test for feature selection is 0.956 and by using Kendall correlation for feature selection is 0.955. Precession of logistic regression by using Pearson correlation method for feature selection is 0.9365, by using chi squared test for feature selection is 0.936 and by using Kendall correlation for feature selection is 0.936. Recall of logistic regression by using Pearson correlation method for feature selection is 1, by using chi squared test for feature selection is 0.998 and by using Kendall correlation for feature selection is 0.997.

Accuracy of neural network by using Pearson correlation method for feature selection is 0.9431, by using chi squared test for feature selection is 0.942 and by using Kendall correlation for feature selection is 0.941. F1 score of neural network by using Pearson correlation method for feature selection is 0.9601, by using chi squared test for feature selection is 0.959 and by using Kendall correlation for feature selection is 0.9579. Precession of neural network by using Pearson correlation method for feature selection is 0.9430, by using chi squared test for feature selection is 0.9407 and by using Kendall correlation for feature selection is 0.9334 . Recall of neural network by using Pearson correlation method for feature selection is 1, by using chi squared test for feature selection is 0.999 and by using Kendall correlation for feature selection is 0.999.

Table.2: Comparison of various feature selection methods accuracy, F1 score, precession and recall .

Classification technique	Feature scoring method	Accuracy	F1 score	precession	Recall
Logistic regression	**Pearson correlation**	0.941	0.9567	0.9365	1
	Chi squared	0.94	0.956	0.936	0.998
	Kendall correlation	0.939	0.955	0.936	0.997
Neural network	**Pearson correlation**	0.9431	0.9601	0.9430	1
	Chi squared	0.942	0.959	0.9407	0.999
	Kendall correlation	0.941	0.9579	0.9334	0.999
Decision tree	**Pearson correlation**	0.939	0.9557	0.933	0.9989
	Chi squared	0.937	0.9546	0.9324	0.9958
	Kendall correlation	0.936	0.9522	0.9315	0.9954
Support vector machine	**Pearson correlation**	0.886	0.8885	0.904	0.931
	Chi squared	0.882	0.8866	0.8877	0.914
	Kendall correlation	0.88	0.882	0.899	0.8949

Accuracy of decision tree by using Pearson correlation method for feature selection is 0.939, by using chi squared test for feature selection is 0.936 F1 score of decision tree by using Pearson correlation method for feature selection is 0.9557, by using chi squared test for feature selection is 0.9546 and by using Kendall correlation for feature selection is 0.9522. Precession of decision tree by using Pearson correlation method for feature selection is 0.933, by using chi squared test for feature selection is 0.9324 and by using Kendall correlation for feature selection is 0.9315. Recall of decision tree by using Pearson correlation method for feature selection is 0.9989, by using chi squared test for feature selection is 0.9958 and by using Kendall correlation for feature selection is 0.9954.

Accuracy of support vector machine by using Pearson correlation method for feature selection is 0.886, by using chi squared test for feature selection is 0.882 and by using Kendall correlation for feature selection is 0.88. F1 score of support vector machine by using Pearson correlation method for feature selection is 0.8885, by using chi squared test for feature selection is 0.8866 and by using Kendall correlation for feature selection is 0.882. Precession of support vector machine by using Pearson correlation method for feature selection is 0.904, by using chi squared test for feature selection is 0.8877 and by using Kendall correlation for feature selection is 0.899. Recall of support vector machine by using Pearson correlation method for feature selection is 0.931, by using chi squared test for feature selection is 0.914 and by using Kendall correlation for feature selection is 0.894

From this table we conclude that from all feature selection methods Pearson correlation method finds best result for all classification techniques. Now we compare various machine learning techniques their accuracy, f1 score, precession, and recall.

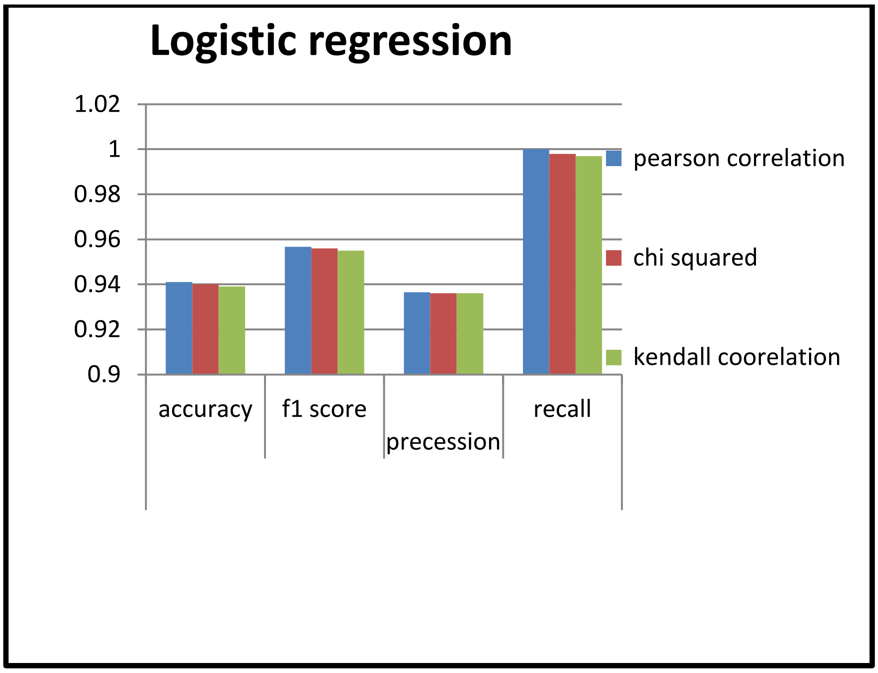

Fig.2 : comparison of accuracy, f1 score, precession and recall of various feature selection methods of logistic regression a machine learning technique.

Following graph compares the accuracy, F1 score, Precession, Recall of logistic regression by using feature selection methods like Pearson correlation method, chi squared test and Kendall correlation and finds the best result.

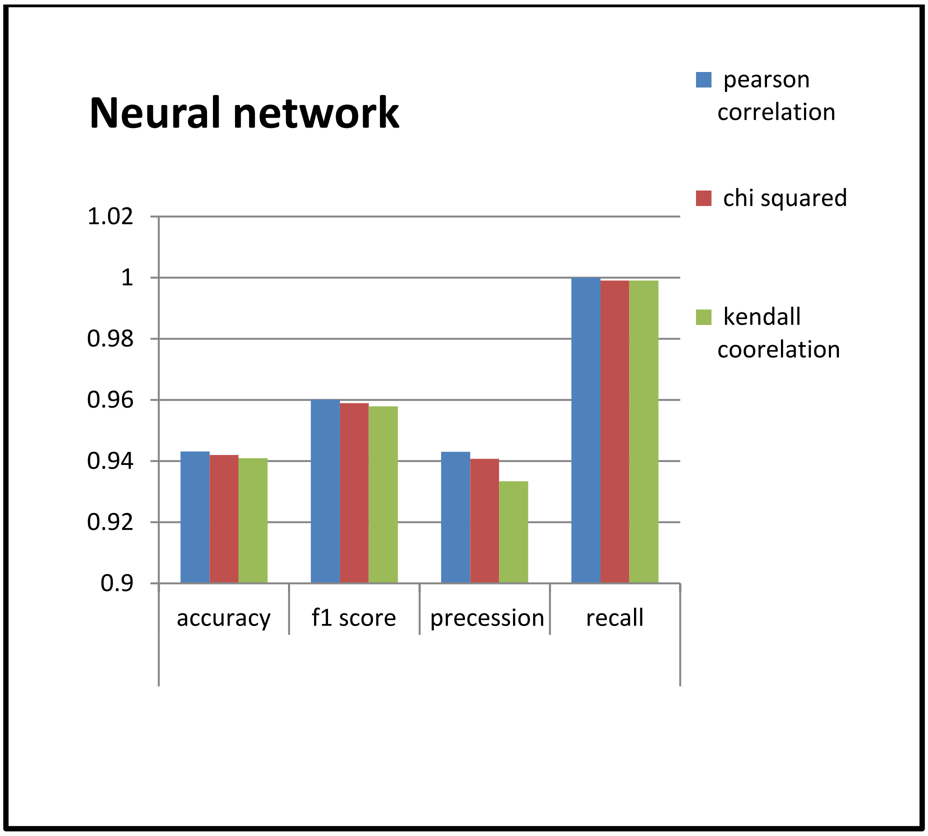

Fig 3 : comparison of accuracy, f1 score ,precession an recall of various feature selection methods on neural network machine learning techniques

Following graph compares the accuracy, F1 score, Precession, Recall of neural network by using feature selection methods like Pearson correlation method, chi squared test and Kendall correlation and finds the best result.

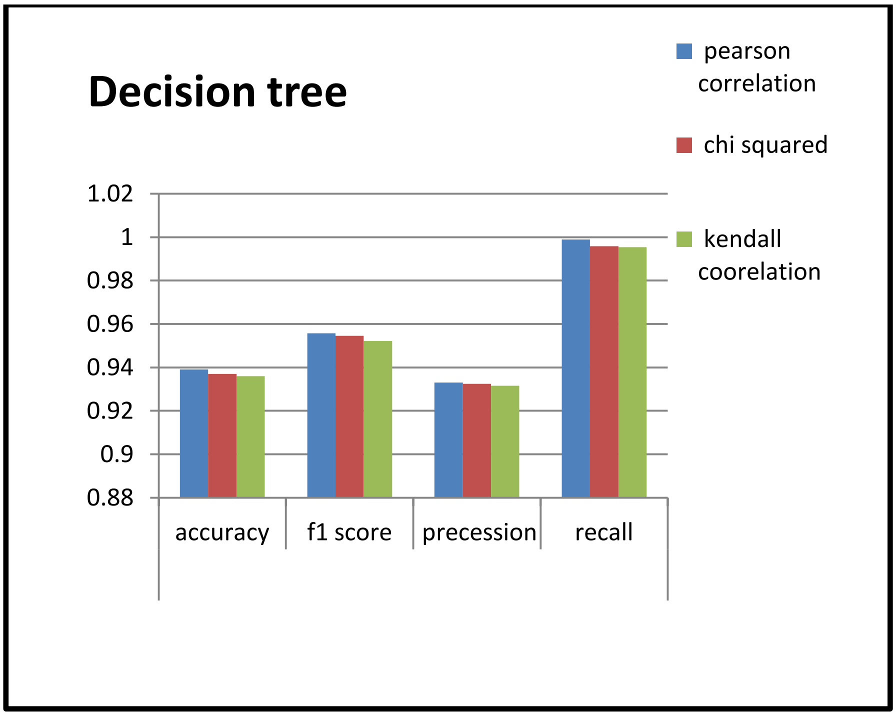

Fig 4 : comparison of accuracy, f1 score, precession and recall of various feature selection methods on decision tree machine learning techniques .

Following graph compares the accuracy, F1 score, Precession, Recall of decision tree by using feature selection methods like Pearson correlation method; chi squared test and Kendall correlation and finds the best result.

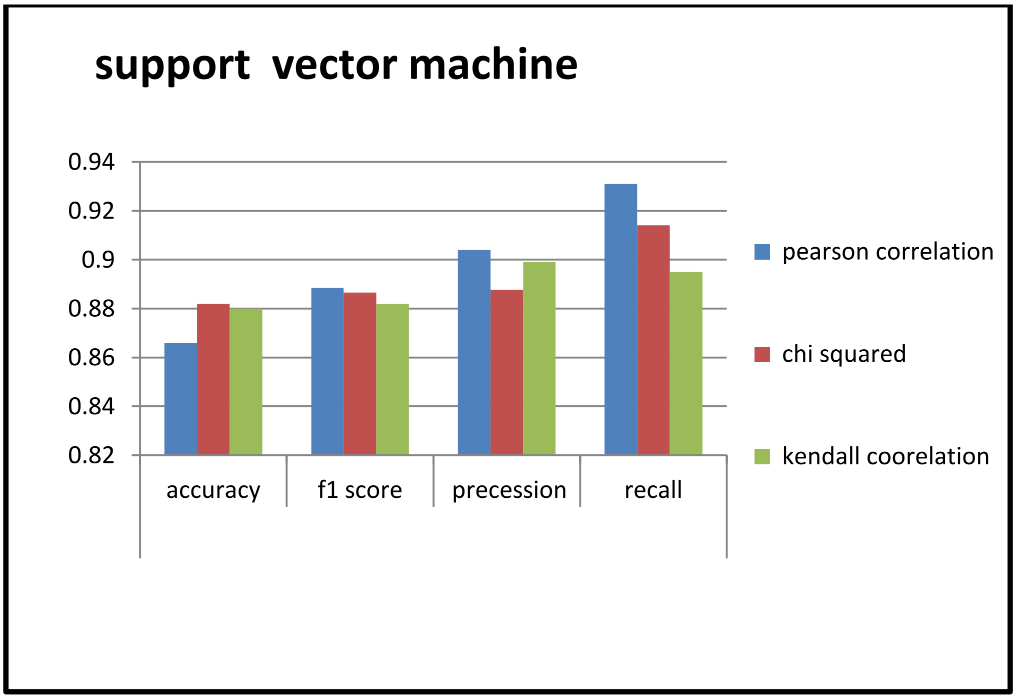

Fig 5 : comparison of accuracy, f1 score, precession and recall of various feature selection methods on neural network machine learning techniques.

Following graph compares the accuracy, F1 score, Precession, Recall of support vector machine by using feature selection methods like Pearson correlation method, chi squared test and Kendall correlation and finds the best result.

Table.3 :various machine learning technique and comparison of their accuracy , F1 score , Precision and Recall

Classificastion technique	Accuracy	F1 Score	Precision	Recall
Logistic regression	0.941	0.9567	0.9365	1
Neural network	0.9431	0.9501	0.9430	1
Decision tree	0.939	0.9557	0.933	0.9989
Support vector machine	0.886	0.8885	0.904	0.931

The results show that the neural network find the highest accuracy, precision recall and f1 score as of 0.9531, and the support vector machine classification obtained the worst result as of 0.886 .

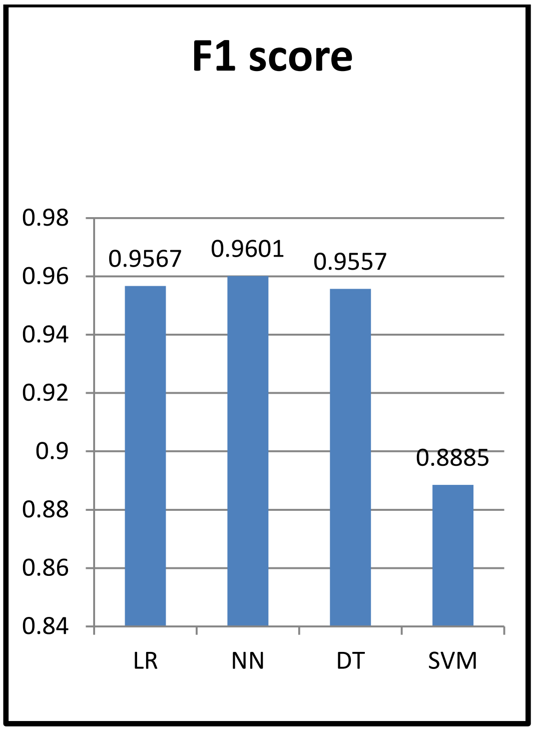

Fig 6 : comparison of f1 score of various techniques

Fig shows the Comparison of F1 score of various machine learning technique like logistic regression ,neural network ,decision tree ,and support vector machine.

F1 score of logistic regression technique is 0.9567 ,neural network is 0.9601 , decision tree is 0.9557 and support vector machine is 0.8885.

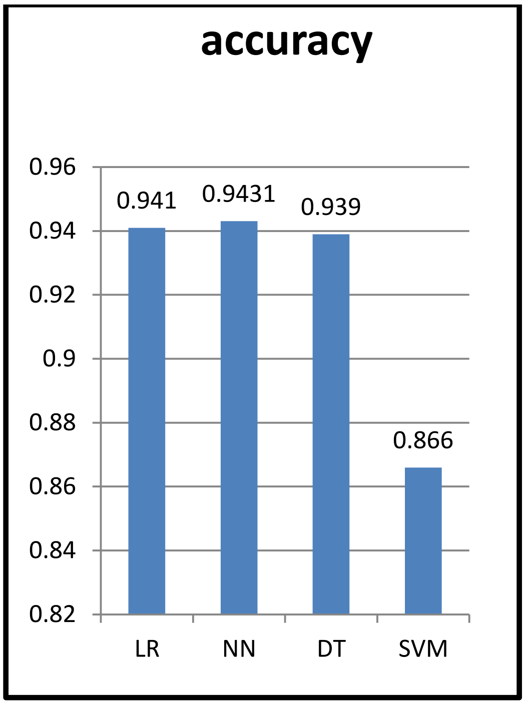

Fig 7 : comparison of accuracy of various techniques

Fig shows the Comparison of accuracy of various machine learning technique like logistic regression, neural network, decision tree, and support vector machine.

Accuracy of logistic regression technique is 0.941, neural network is 0.9431 , decision tree is 0.939 and support vector machine is 0.866.

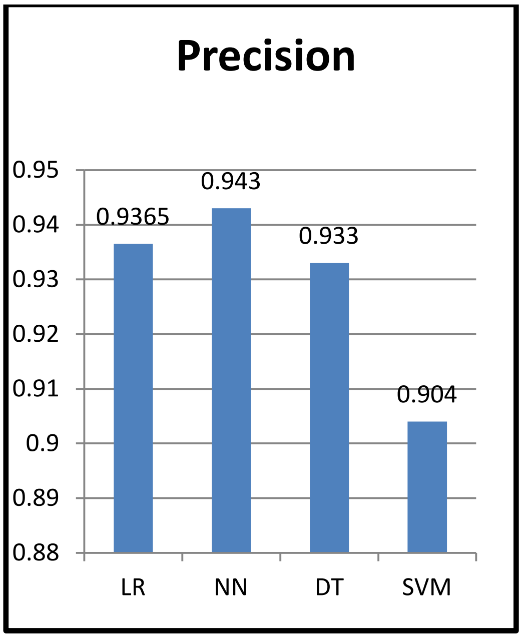

Fig.8 : comparison of Precision of various machine learning techniques

Fig shows the Comparison of precession of various machine learning technique like logistic regression ,neural network ,decision tree ,and support vector machine.

Accuracy of logistic regression technique is 0.9365 ,neural network is 0.943 , decision tree is 0.933 and support vector machine is 0.904.

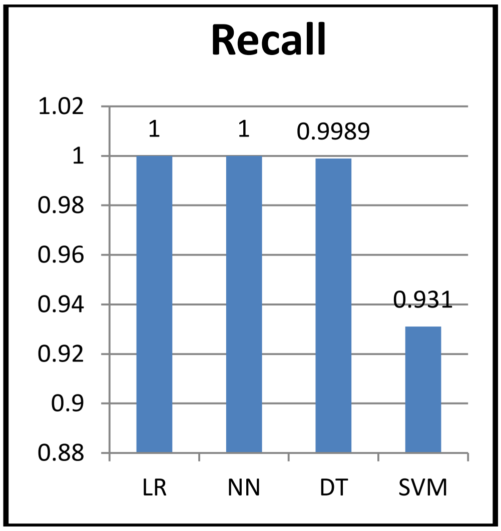

Fig.9 : comparison of Recall of various machine learning techniques

Fig shows the Comparison of Recall of various machine learning technique like logistic regression, neural network, decision tree, and support vector machine.

Recall of logistic regression technique is 1, neural network is 1, decision tree is 0.9989 and support vector machine is 0.931.

5. CONCLUSIONS

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 we 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)is proposed to detect spam

6. REFERENCES

[1] Almomani, Ammar, B. B. Gupta, SamerAtawneh, A. Meulenberg, and Eman Almomani. exercises 15, no. 4 pp. 2070-2090,2013 *"A review of phishing email separating procedures."* IEEE correspondences overviews and instructional .

[2] Gansterer, W. N., and Pölz, D., pp. 449-460,2009 "*Email characterization for phishing protection."* In Advances in Information Retrieval,.Springer Berlin Heidelberg .

[3] McGregor, Anthony, Mark Hall, Perry Lorier, and James Brunskill. pp. 205-214, 2004 "Stream bunching utilizing machine learning strategies." *In International Workshop on Passive and Active Network Measurement ,Springer, Berlin, Heidelberg, .*

[4] Read, Jonathon. , pp. 43-48, 2005 "Utilizing emojis to lessen reliance in machine learning methods for slant characterization." *In Proceedings of the ACL understudy investigate workshop,Relationship for Computational Linguistics*.

[5] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. *building* 160 pp 3-24.,2007, "Regulated machine taking in: An audit of arrangement strategies." *Emerging man-made reasoning applications in PC.*

[6] Rathi, M., &amp; Pareek, V. 2013 "Spam Mail Detection through Data Mining-A Comparative Performance Analysis". *International Journal of Modern Education and Computer Science*,(12).

[7] Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. , pp. 60-69 , 2007 "An examination of machine learning procedures for phishing identification." *In Proceedings of the counter phishing working gatherings second yearly eCrime analysts summit.*

[8] Sommer, Robin, and Vern Paxson. , pp. 305-316 , 2010 "Outside the shut world: On utilizing machine learning for arrange interruption location." *IEEE.*

[9] Kolari, Pranam, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. vol. 6, pp. 1351-1356. 2006 "Distinguishing spam writes: A machine learning approach." *In AAAI.*

[10] Crawford, Michael, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. no. 1: 23,2015 *"Overview of audit spam location utilizing machine learning systems." Journal of Big Data .*

[11] Wang, Alex Hai. , pp. 335-342, 2010 "Identifying spam bots in online long range interpersonal communication locales: a machine learning approach." *In IFIP Annual Conference on Data and Applications Security and Privacy,. Springer, Berlin, Heidelberg.*

[12] Castillo, Carlos, Debora Donato, Aristides Gionis,

Vanessa Murdock, and FabrizioSilvestri. pp. 423-430, 2007 "Know your neighbors: Web spam discovery utilizing the web topology." *In Proceedings of the 30th yearly worldwide ACM SIGIR gathering on Research and advancement in data recovery*.

[13] Benevenuto, Fabricio, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. vol. 6, pp. 12, 2010 "Recognizing spammers on twitter." *In Collaboration, electronic informing, hostile to manhandle and spam meeting (CEAS)*.

[14] Sasaki, Minoru, and Hiroyuki Shinnou. Vol. 4 , 2005 "Spam location utilizing content bunching." *In Cyberworlds,2005.worldwide meeting , IEEE* .

[15] Garera, Sujata, Niels Provos, Monica Chew, and Aviel D. Rubin. *malcode*, pp. 1-8, 2007 "A structure for discovery and estimation of phishing assaults." *In Proceedings of the ACM workshop on Recurring* .

[16] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Vol.10, pp. 79-86, 2002 "Thumbs up: slant arrangement utilizing machine learning systems." *In Proceedings of the ACL-02 meeting on Empirical techniques in normal dialect preparing*.

[17] Witten, Ian H., Eibe Frank, Mark A. Lobby, and Christopher J. Buddy. 2016. " Information Mining: Practical machine learning devices and systems."

[18] Kumar, R. K., Poonkuzhali, G., and Sudhakar, P. Vol. 1, pp. 14-16,march-2012 *"Similar investigation on email spam classifier utilizing information mining procedures"*. In Proceedings of the International Multi Conference of Engineers and Computer Scientist.

[19] Jyoti Chhikara, CSE Dept, PDMCEW India: Volume 3, Issue 5, May 2013." International Journal of Advanced Research in Computer Science and Software Engineering"