IRSTI 28.23.27 https://doi.org/10.26577/JPCSIT.2023.v1.i2.02

M. Medelbekov , M. Nurtas* , A. Altaibek

International Information Technology University, Kazakhstan, Almaty
*e-mail: maratnurtas@gmail.com

MACHINE LEARNING METHODS FOR PHISHING ATTACKS

Abstract. The basis of cybersecurity is an understanding of the mechanisms of social engineering. This increases the effectiveness in combating this type of manipulation. One of them is phishing. Phishing attacks actively exploit the human factor to collect credentials or distribute malware. Phishing websites are visually similar to real websites. Along with the development of technology, phishing methods have also evolved. Machine learning(ML) has been effectively used to identify and avoid phishing. The reason of this consider is to survey ML methods and the comes about of previous thinks about on the avoidance of phishing attacks. As well as our claim investigation and execution of a model for recognizing phishing sites. The efficiency of the demonstrate is moved forward by combining connected parameters. 5 calculations were utilized to prepare the show: Logistic Regression, Random Forest, Support Vector Machine(SVM), K-nearest neighbors algorithm(KNN) and KNN k-Fold Cross Validation.

Key words: Phishing, Machine Learning (ML), Cybersecurity, Logistic Regression, Random Forest, KNN, SVM.

Introduction

Identity theft is still one of the foremost unsafe crimes for Internet users [1]. Attackers can impersonate another person not only for the purpose of theft, but also use a person's personal information to commit other crimes [2].

According to Symantec, the leading information security and antivirus software company, there is one phishing email for every 2,000 emails [3]. In 2020, Google stated that systems register 25 billion spam pages every day [4]. Phishing attacks are not uncommon.

- Phishing Sources

Systematic phishing incidents have originated on America Online (AOL) since 1995 [5]. Since then, different types of attacks have appeared, differing by source:

- Email of the foremost dangerous phishing is the foremost common sort of phishing that almost everyone has encountered. Usually it goes straight to the "spam" section in the mail.
- Smishing is an attack by sending illegitimate SMS messages.
- Vishing – phishing using mobile calls. Vishing calls send fictitious notifications from banks.
- Evil Twin Phishing is an attack by replacing a legitimate Wi-Fi network with a copy network. As a result, a person connects to a malicious network and, at the authorization stage, enters information that's sent to the attacker.
- Phishing in social networks, search engines, etc.

There are many sources of phishing attacks: from SMS and social networks to email and search engines [6].

- Phishing Targets

Phishing is carried out with the intention of stealing personal data: logins and passwords in instant messengers, numbers and SMS codes of bank cards, passport data, etc [7]. An attacker can get rich by selling information, blackmailing people, and sometimes entire companies. Phishing subtypes are distinguished for the latter:

- Spear Phishing and an attack aimed at employees of a particular company;
- Whaling – phishing of the leaders of a particular organization.

The consequences of such attacks vary. For example, in November 2020, a whaling attack was carried out against the co-founder of the Australian hedge fund Levitas Capital [8]. Fraudsters introduced malware into the corporate network of the fund under the guise of Zoom. The virus almost led to the withdrawal of 8.7 million US dollars to the accounts of scammers. As a result, the organizers of the attack received 800 thousand dollars, and the reputational damage of Levitas Capital led to the closure of the company.

To conclude the above, social engineering techniques are used by phishers to steal personal information by sending emails or messages that link to malware or websites. A phishing attack can be massive with random victims or targeted when the attack is directed at a specific person. Phishing usually

looks like a real email, so it's hard to identify even with reliable security software.

Materials and Methods

All work comprises of the taking after steps:

- Load the data and prepare it for use.
- Do EDA analysis and examine the data.
- Train the model and see the result.

There are algorithms description and its applications used in this work.

Logistic regression. Logistic regression is a useful ML algorithm for binary classification problems, and it can be applied in various domains, including phishing detection, fraud detection, and medical diagnosis. Within the setting of phishing, logistic regression can be utilized to recognize whether an mail or site is likely to be a phishing assault or not. The model is trained on a dataset of examples of phishing and legitimate websites/emails, where each example is labeled with its corresponding class (phishing or not phishing).

The logistic regression model learns the relationship between the input features (such as the URL, sender's email address, content of the email) and the corresponding output labels (phishing or not phishing). Once trained, the model can be used to predict the likelihood of a new email or website being a phishing attack, based on the input features.

In practical applications, logistic regression can be used in phishing detection systems to automatically classify emails or websites as phishing or not phishing, based on the model's predictions. The framework can at that point alarm the client or take other activities to avoid the client from falling casualty to a phishing assault.

The success of logistic regression depends on the quality and quantity of the input data, the appropriate choice of hyperparameters, and the effective evaluation and tuning of the model's performance.

Random Forest. Random forest can be applied in phishing detection by training the model on a dataset of examples of phishing and legitimate emails/websites, where each example is labeled with its corresponding class (phishing or not phishing). The input features can include information such as the URL, sender's email address, content of the email, and other relevant information.

The main part of this algorithm is building multiple decision trees on random subsets of the data. Each decision tree is trained on a random sample of the input features and output labels. When building each decision tree, the algorithm splits the nodes based on the input features that best separate the output labels. The splitting process continues until the tree is fully grown, or a stopping criterion is met. Once all the decision trees have been built, the algorithm combines their predictions to produce a final output. This is done by aggregating the predictions from each decision tree and selecting the class with the most votes.

Random Forest model can be used to predict the likelihood of a new email or website being a phishing attack. The model can be integrated into a phishing detection system that automatically classifies emails or websites as phishing or not phishing based on the model's predictions. This can help prevent users from falling victim to phishing attacks and improve overall cybersecurity.

Support Vector Machine. SVM can be applied in phishing detection by training the model on a dataset of examples of phishing and legitimate emails/websites, where each example is labeled with its corresponding class (phishing or not phishing). The input features can include information such as the URL, sender's email address, content of the email, and other relevant information.

The first step of SVM algorithm is to prepare the dataset for training the model. This includes cleaning the data, handling missing values, and converting categorical variables to numerical format. At that point the user chooses a set of highlights that are pertinent to the classification issue. This makes a difference to diminish the dimensionality of the dataset and progress the execution of the demonstrate. SVM tries to discover the hyperplane that maximizes the edge between the classes of information. The margin is the remove between the hyperplane and the closest information focuses from each lesson. The hyperplane that maximizes the margin is the one that has the best generalization performance on new, unseen data. If the data is not linearly separable, SVM uses a technique called the kernel trick to transform the input data into a higher-dimensional space where it is linearly separable.

Once trained, the SVM model can be used to predict the likelihood of a new email or website being a phishing attack. The model can be integrated into a phishing detection system that automatically classifies emails or websites as phishing or not phishing based on the model's predictions. This can help prevent users from falling victim to phishing attacks and improve overall cybersecurity.

K-Nearest Neighbors. KNN can be applied in phishing detection by training the model on a dataset of examples of phishing and legitimate emails/

websites, where each example is labeled with its corresponding class (phishing or not phishing). The input features can include information such as the URL, sender's email address, content of the email, and other relevant information.

KNN algorithm principles:

1) Choosing the value of K: The user selects a value for K, which is the number of nearest neighbors that will be used to classify or predict the output label of a new observation.

2) Calculating distances: The algorithm calculates the distance between the new observation and all the existing observations in the training dataset. The distance can be calculated using various metrics, such as Euclidean, Manhattan, or Minkowski distance.

3) Identifying K nearest neighbors: The algorithm identifies the K nearest neighbors to the new observation based on the calculated distances.

4) Classifying or predicting the output label: Once the K nearest neighbors have been identified, the algorithm uses a majority voting scheme to classify or predict the output label of the new observation. For classification problems, the output label is the class with the most votes among the K nearest neighbors. For regression issues, the yield name is the normal of the output values among the K closest neighbors.

The KNN model can be used to predict the likelihood of a new email or website being a phishing attack. The model can be integrated into a phishing detection system that automatically classifies emails or websites as phishing or not phishing based on the model's predictions. This can help prevent users from falling victim to phishing attacks and improve overall cybersecurity.

Dataset overview

The dataset consists of different Phishing Websites Features.

The features presented in this dataset are an almost ideal dataset for training models to

distinguish phishing locales. They have become the basis of many studies, proving their reliability and effectiveness in forecasting.

When determining a website's URL as legitimate or phishing, there are various features that need to be taken into account, including many parameters.

Components for detecting and classifying phishing websites:

1. Address Bar based Features
2. Abnormal Based Features
3. HTML and JavaScript Based Features
4. Domain Based Features

Address Bar based Features are the following: Using the IP Address**,** Long URL to Hide the Suspicious Part, Using URL Shortening Services "TinyURL", URL's having "@" Symbol, Redirecting using "//", Adding Prefix or Suffix Separated by (-) to the Domain, Sub Domain and Multi Sub Domains, HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer), Domain Registration Length, Favicon, Using Non-Standard Port and The Existence of "HTTPS" Token in the Domain Part of the URL. Table 1 presents common ports to be checked for avoiding phishing attacks.

Table 1 – Common ports to be checked

#	Service	Meaning	Preferred Status
1	FTP	Transfer files from one host to another	Close
2	SSH	Secure File Transfer Protocol	Close
3	Telnet	Provide a bidirectional interactive text-oriented coomunication	Close
4	HTTP	Hyper Test Transfer Protocol	Close
5	HTTPS	Hypertext Transfer Protocol Secured	Close
6	SMB	Providing shared access to files, printers, serial ports	Close
7	MSSQL	Store and retrieve data as requested by other software applications	Close
8	ORACLE	Access oracle database from web	Close
9	MySQL	Access MySQL database from web	Close
10	Remote Desktop	Allow remote access and remote collaboration	Close

Abnormal Based Features include Request URL, URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Server Form Handler (SFH), Submitting Information to Email, and Abnormal URL.

Website Forwarding, Status Bar Customization, Disabling Right Click, Using Pop-up Window and IFrame Redirection are the HTML and JavaScript based Features.

Finally, Domain based Features include, Age of Domain**,** DNS Record, Website Traffic, PageRank, Google Index, Number of Links Pointing to Page, Statistical-Reports Based Feature.

The data was presented as encoded values -1, 0 and 1. We replaced the values with strings in accordance with the documentation. That is, "-1" means legitimate, "0" means suspicious, "1" means phishing. The original dataset with encoded data has been saved in a copy.

The dataset consists of 30 columns and 11055 rows. There are no missing values. Next comes the visualization part.

In machine learning, visualizing the distribution of each column in a dataset can provide valuable insights into the underlying data and help identify potential issues that may need to be addressed during preprocessing or modeling.

When we visualize the distribution of a column, we are essentially plotting the frequency of each value in the column, typically using a histogram or a density plot. This allows us to see how the values are spread out across the range of possible values and whether there are any outliers or unusual patterns.

Some of the things that we can learn from visualizing each column distribution include:

- Whether the data is normally distributed or skewed: A normal distribution is symmetrical and has a bell-shaped curve, while a skewed distribution is asymmetrical and has a tail that is longer on one side than the other.
- • Whether there are any exceptions: Exceptions are information focuses that are essentially distinctive from the other values within the dataset, and can have a expansive affect on the execution of a machine learning show.
- Whether there are any missing values: Missing values can be represented by gaps in the histogram or density plot, and may need to be imputed before the data can be used for modeling.
- Whether there are any categorical variables that need to be encoded: If a column contains categorical variables, we may need to encode them using techniques such as one-hot encoding or label encoding before they can be used for modeling.

Visualizing the distribution of each column in a dataset is an important step in the data preprocessing and exploratory data analysis process in machine learning. It can help us gain a better understanding of the underlying data and make informed decisions about how to preprocess and model the data. Figure 1 presents the distribution of the target variable.

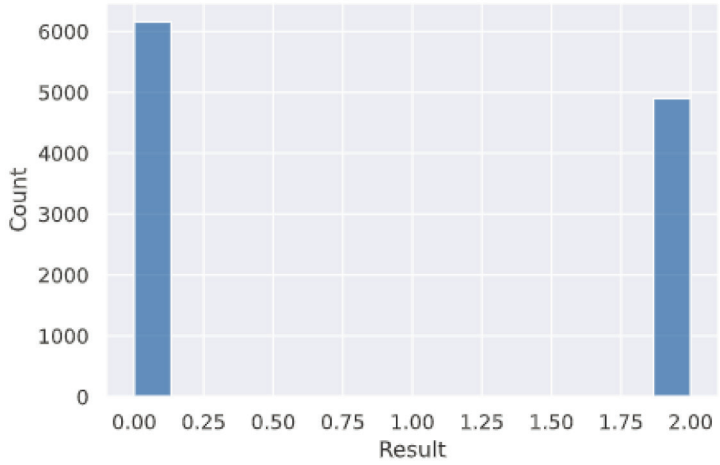

Figure 1 – Target variable distribution plot

A correlation heatmap is a graphical representation of a correlation matrix, which shows how strongly different variables are related to each other. The heatmap is a grid of squares that are color-coded to represent the strength and direction of the correlation between two variables.

To get it a relationship heatmap, it's vital to know what the color coding implies. Regularly, shades of blue show negative relationship (when one variable increments, the other variable diminishes) and shades of ruddy demonstrate positive relationship (when one variable increments, the other variable increments as well). The escalated of the color demonstrates the quality of the relationship, with darker colors speaking to more grounded relationships.

When translating a relationship heatmap, it is vital to keep in mind that relationship does not fundamentally infer causation. Just because two variables are strongly correlated does not mean that one causes the other. Additionally, correlation coefficients can be influenced by outliers or other factors, so it is important to look at the data underlying the heatmap and not rely solely on the color coding. The correlation heatmap of the data is on the Figure 2.

As we can see the highest correlation is 94% between "popUpWindow" and "Favicon" parameters. Then is 84% between "double_slash_redirecting" and "Shortining_Service".

Figure 3 demonstrates top 3 variables with high correlation with respect to the target variable "Result" are "Prefix_Suffix", "Request_URL" and "Google_Index".

These observations give a big picture of the dependence and influence of parameters on each other. Using this matrix, it is possible to look for more efficient approaches to combining the parameters of the functions of phishing websites.

In an algorithmic approach, a machine learning model is some parameterized function f(·, θ)/ X --> Y. For example, it can be linear regression, a neural network, an ensemble decision tree, or a support vector machine (although the last two models have a variable number of parameters, but this is not important). Such a model directly predicts the value of y. In the probabilistic approach, the model still predicts the number f(x, θ), but now this number is considered not the final prediction of y, but the expectation of a normal distribution with some fixed variance σ. Thus, the conditional distribution p(y|x) is modeled. The probabilistic show essentially formalizes what was inferred casually within the point assess.

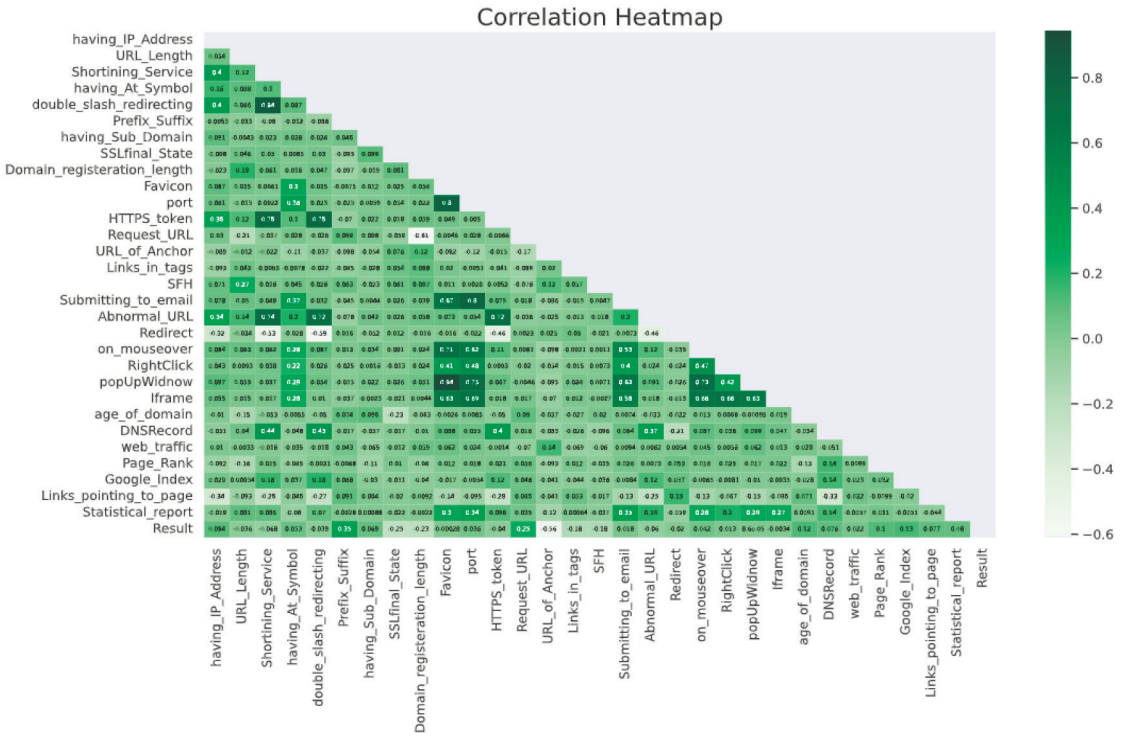

Figure 2 – Correlation heatmap

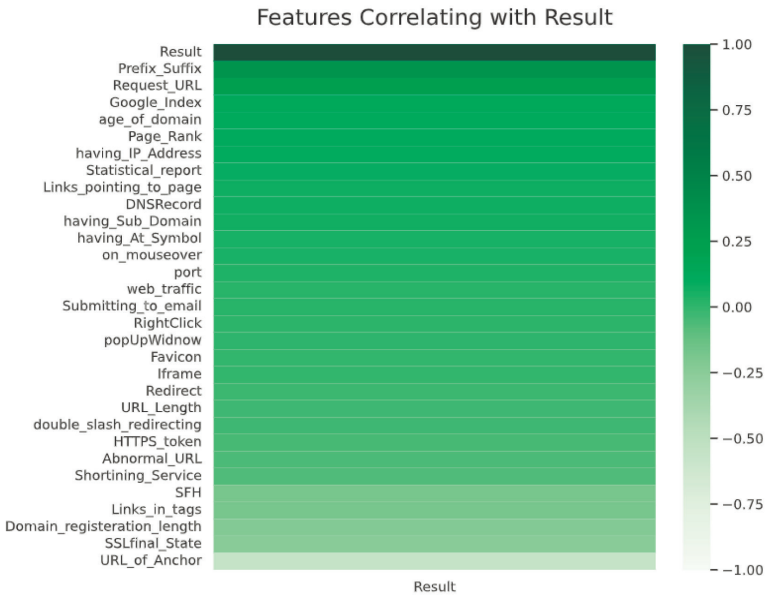

Figure 3 – Correlation heatmap with target variable

Our model, due to the dispersion of the test of normal values on Figure 4, can beat ordinary models prepared with the MSE loss work. Maybe typically due to the reality that the retraining of the model on exceptions is decreased, since a tall esteem of vulnerability is anticipated in regions with visit exceptions, which mellows the expectation blunder of the scientific desire in these ranges, lessening the multiplier at the primary term.

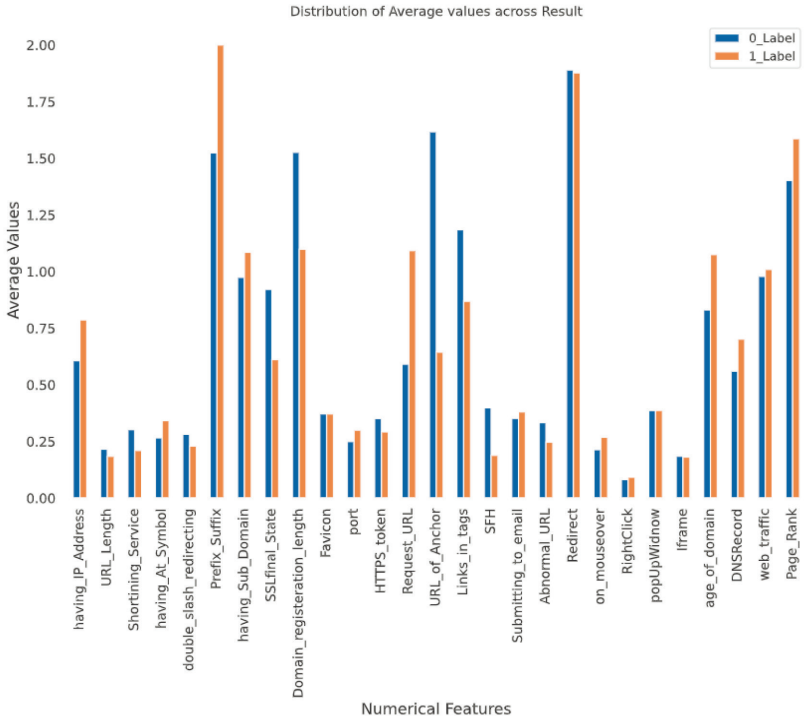

Figure 4 – Average values distribution accross Result target variable

Literature review

The phishing attack mechanism targets human, not systemic, vulnerability. This makes the task of detecting an attack more difficult. Separation into phishing or safe sources is suitable for the classification task, so the use of ML is relevant here.

ML is a field of artificial intelligence, and it trains a computer to extract information from data, identify certain patterns and make decisions according to the chosen model.

Currently, the following ML algorithms are widely used due to their performance and high accuracy: Decision Tree, Random Forest, K-Means Clustering, Naive Bayes, SVM, and Artificial Neural Network (ANN). The Table 1 describes these methods.

R.P. Ferreira with other researchers used ANN Multilayer Perceptron (ANN-MP) to calculate phishing attacks [9]. ANNs resemble the structure of the brain and imitate some functions of human behavior (abstraction, generalization, learning). In their work, they correctly classified websites with phishing characteristics with an accuracy of 87.61%. During the test phase, the accuracy of ANN-MLP was 98.23%. The paper also provides a comparative analysis with other works in which artificial intelligence methods were used. The ANN-MLP developed by the authors showed one of the best results. In future studies, it was planned to change the order of attributes to find the best groups, increase the database for training and testing, and improve the performance of solutions to the classification problem. The Figure 5 demonstrates basic elements of an artificial neuron.

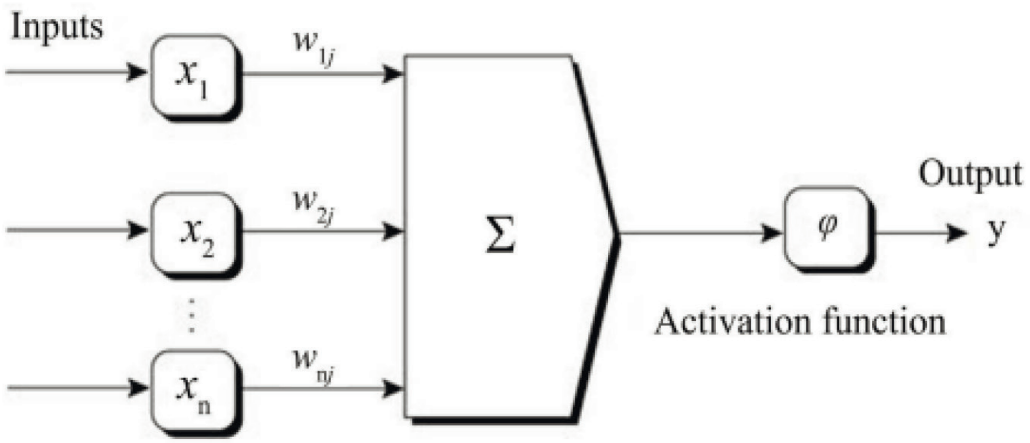

Figure 5 – Basic elements of an artificial neuron [10]

N.V. Puri et al. optimization of the black list method is presented [11]. When developing the approach, K-Means and Naïve Bayes algorithms were used, which checked the sites for their presence in the blacklist, as well as their behavior. The URL features are first extracted, then the K-Means algorithm checks the page for blacklisting. If the page is suspicious, then it passes an additional check through the Naive Bayes algorithm. Figure 6 shows the mechanism of the method in more detail.

In the future, the authors want to implement anti-phishing tools in various environments.

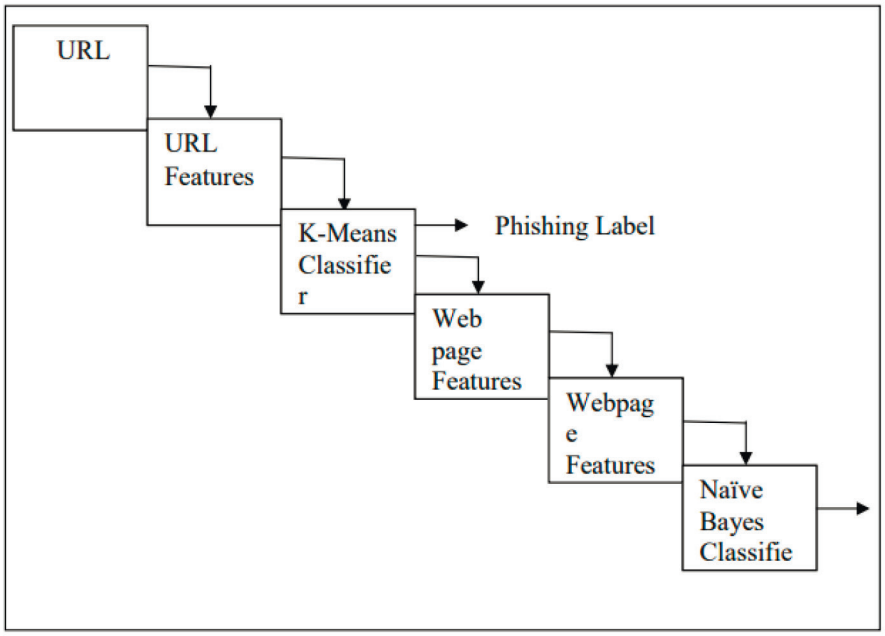

Figure 6 – System Architecture [11]

Analysts M. N. Alam et al. created their claim show utilizing arbitrary timberland (RF) and choice tree (DT) calculations [12]. The dataset for preparing was taken from Kaggle. Vital component investigation (PCA) was utilized to analyze the highlights of the information. The greatest precision of 97% was appeared by the arbitrary timberland calculation.

However, the research was not limited to using conventional machine learning models. The work of Pandey A. et al. proposes to combine two algorithms, random forest and support vector machine (SVM), thereby progressing the comes about of the model [13]. After training, the hybrid model predicted with 94% accuracy (SVM accuracy was 90% and random forest accuracy was 92.96%).

Adwan Yasin and Abdelmunem Abuhasan introduce the concept of weighting phishing terms [14]. It estimates the weight of phishing terms in each email. In the preprocessing phase, the definition of text roots and WordNet ontology was used to enrich the model with word synonyms. The model contained knowledge discovery procedures using Random Forest, J48 Decision Tree, SVM, Multi-Layer Perceptron (MLP), and Naive Bayes. To avoid overfitting training and test data, a 10-fold cross-validation method was used. The accuracy of the model using J48 was 98.4%, and the accuracy of the random forest algorithm model was 99.1%. The figures are among the highest to date. In the future, the authors are going to improve the developed model to obtain better results.

A new approach to the phishing classification problem using K-means algorithms was presented by Vidya Mhaske-Dhamdhere and Sandeep [15]. Email was analyzed in real time using the K-means algorithm. After checking 160 e-mails from computer science students, the results were as follows: True positive for legitimate email was 67% and phishing 80%. True negative for a legitimate 30% and 20% for a phishing email.

Author in Jaypee Institute of Information Technology describe an approach of detecting phishing websites via 5 ML algorithms [16]. The paper presented methods for evaluating the prediction of each model. Website attributes were extracted using Python, and performance evaluation is performed using the R language. The best result was given by the Random Forest algorithm. Its accuracy were 98.4%, recall 98.59% and precision 97.70%.

Deshpande A. et al. used a combination of URL lexical features and other features such as host [17]. Thus, it turned out that this is the most effective approach to combat phishing. The accuracy of the Random Forest algorithm averaged about 97%, and for the Decision Tree – 95%. For future improve-

ments, the researchers intend to improve the accuracy of the models through better feature extraction, as well as build a phishing detection system as a scalable web service with online training.

Phishing attacks can also be classified as a classification problem. This was suggested in their work by Y. Sönmez et al [18]. Their classification model includes site feature extraction and website classification. From the dataset of the UCI Irvine ML repository, they took 30 features were taken with well-defined rules for extracting phishing features. Different machine learning strategies were utilized for classification, such as bolster vector machines (SVM), credulous bayes (NB), and extraordinary learning machine (ELM). Compared to SVM and NB, Extreme Learning Machine (ELM) achieved 95.34% accuracy in six uses of six activation functions. The results were obtained using MATLAB.

Desai et al. approached the problem on a larger scale in their work. They created a special extension for Google Chrome. Using machine learning algorithms, it detects the content of phishing sites [19]. The data was taken from the UCI machine learning store, from which 22 features were extracted. Then we compared the accuracy, recall, f1-speed of 3 algorithms: kNN, SVM and Random Forest. The best result was shown by Random Forest. The Chrome extension was implemented using HTML, JavaScript, CSS, and Python.

Table 2 – Comparative analysis of the machine learning methods for phishing detection

#	Study	Method	Advantages	Disadvantage
1	2	3	4	5
1	R. P. Ferreira et al., Artificial Neural Network for Websites Classification with Phishing Characteristics, (2018)	ANN	In this approach, ANN allows you to specify additional parameters, such as an attribute and a training type	The comes about of the classification handle straightforwardly depend on the arrange of the information qualities
2	M. N. Badadhe et al., An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms With Results (2014)	k-means	In the feature space, this approach minimizes the clustering error	This approach does not define the website as a phishing site, but considers it as "probably phishing", i.e. does not solve the classification problem
3	M. N. Alam et al., Phishing Attacks Detection using Machine Learning Approach, (2020)	Random Forest, Decision Tree	1. Especially in non-linear problem, RF has better prediction accuracy and performance. 2. With enough trees, RF will not fit the model and avoid overfitting. 3. RF is easy to implement and interpret	1. A huge number of trees may not be relevant for real-time forecasts. 2. RF complicates the interpretation of model relationships. 3. RF is sensitive to small changes in parameter value
4	Pandey A. et al., Identification of phishing attack in websites using random forest-svm hybrid model, (2018)	SVM, Random Forest	When applied to multidimensional data with their minimum volume, SVM is better suited than other algorithms	The classification process is time consuming due to the SVM requirement for a convex combination of kernels
5	A. Yasin and A. Abuhasan, An Intelligent Classification Model For Phishing Email Detection, (2016)	SVM, Decision Tree, Random Forest, Naive Bayes, MLP	1. SVM is better at predicting non-linearly separable data. 2. Decision tree is easier to explain and implement	1. SVM requires a convex combination of cores and the SVM model is difficult to interpret and understand 2. DT does not support online learning and requires the tree to be rebuilt each time new samples appear, i.e. increases processing time 3. NB has low accuracy due to the lack of information about the relationships between features in the samples

1	2	3	4	5
6	Vidya Mhaske-Dhamdhere and Sandeep, A Novel Approach for Phishing Emails Real Time Classification Using K-Means Algorithm, (2018)	k-means	The implementation and execution of the method is simple	If the initialization is incorrect, then the result will be low
7	J. Shad, S et al., A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology, 2018	Decision Tree, Random Forest, Gradient Boosting, Generalized Linear Model, Generalized Additive Model	Website attributes were extracted using Python, and performance evaluation is performed using the R language.	If the initialization is incorrect, then the result will be low
8	Deshpande A. et al., Detection of Phishing Websites using Machine Learning (2021)	Decision Tree, Random Forest	High efficiency of the result due to the combination of parameters for model training	1. A huge number of trees may not be relevant for real-time forecasts. 2. RF complicates the interpretation of model relationships. 3. RF is sensitive to small changes in parameter value
9	Y. Sönmez et al., Phishing web sites features classification based on extreme learning machine, 2018	MATLAB, SVM, Naive Bayes, ELM	The capabilities of ELM can sufficiently reduce amount of training time for one hidden layer of a feedforward neural network	Unlike conventional learning algorithm-based settings, ELM has a requirement for more hidden nodes. This is due to the random definition of input weights and hidden biases [20]-[21].
10	A. Desai et al., Malicious web content detection using machine leaning, 2018.	kNN, SVM, Random Forest	Special Google Chrome extension for detecting phishing sites	The list of declared malicious sites is increasing every day

This section has presented previous relevant and relevant work by various researchers on the topic of detecting phishing websites. The analysis of these works led to the use of machine learning methods in this work. The total number of extracted features will be about 30 items.

Results and Discussion

A confusion grid is used to compare forecasts and reality. It can be a table with 4 different possible combinations of expected and real numbers. The expected values are depicted as positive and negative, while the real values are depicted as true and false. The confusion scheme is often used to assess the accuracy of models in classification issues. But prediction and projection confirmation can be seen as an unusual case of this question, so the confusion matrix is additionally appropriate for measuring the accuracy of predictions. It allows us to investigate the viability not only in subjective terms, but also in quantitative terms.

Table 3 – Confusion Matrix description

Prediction	Positive	Negative
	+	-
+	True Positive (the figure coincided with reality, the result was positive, as anticipated by the ML show)	False Positive (type 1 error, ML model predicted a a positive result, but in fact it is negative)
-	False Negative (sort 2 mistake – ML-model anticipated a negative result, but in truth it is positive)	True Negative (result was negative, ML prediction matched reality)

Comparing all the algorithms, Random Forest has the best Result. Figure 7 represents the confusion matrix of this algorithm.

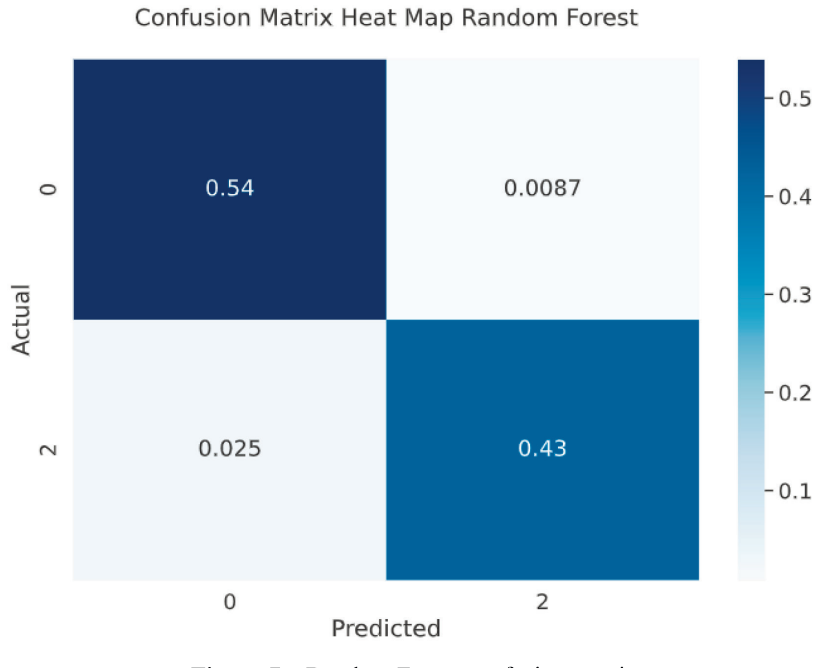

Figure 7 – Random Forest confusion matrix

From a mathematical point of view, the accuracy of an ML model can be assessed using the following metrics:

- Accuracy – how many total results were predicted correctly;
- «Error rate;
- Recall – how many true outcomes were predicted correctly;
- An F-measure that allows you to compare 2 models while evaluating recall and accuracy at the same time. Here, the harmonic mean is utilized rather than the arithmetic mean, smoothing out the calculations by eliminating extreme values.

In quantitative terms, it will look like this:

- P is the number of true results, P = TP + FN;
- N is the number of false positives, N = TN + FP

Table 4 shows the results of model training by different algorithms. All 5 models performed well, but Logistic Regression stings the least at 84 percent. The highest result was obtained by the Random Forest algorithm. It amounted to 96.7%.

Table 4 – Prediction results

Method	Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.84	0.84	0837	0.822	0.834
Random Forest	0.967	0.967	0.90	0.946	0.963
SVM	0.965	0.965			
KNN	0.926	0.926	0.962	0.87	0.914
KNN k-Fold Cross Validation	0.948	0.948	0.90	0.903	0.94

Conclusion

This paper concludes an overview of phishing detection studies. Phishing detection can be attributed to a classification problem, for which ML algorithms are utilized. This review looked at models using Naive Bayes, SVM, Decision Tree, Random Forest, k-means clustering and ANN. Using the above different machine learning methods and according to the accuracy of their final expected results, we have shown from our results which method works more accurately and efficiently.

The best result with an accuracy of 96.7 percent was shown by the Random Forest model. Here we also explained how they can be detected using ML methods using the example of our own developments.

According to the results of the review, the algorithms of SVM, Random Forest and k-means clustering will be applied for our further works.

References

1. Gokhberg, L., Kuznetsova, T. "Strategiya-2020: novye kontury rossiiskoi innovatsionnoi politiki [Strategy 2020: New Outlines of Innovation Policy]." Foresight-Russia 5, no 4 (2011): 8– (In Russian).

2. Ramanathan V., & Wechsler H. (2012). phishGILLNET – phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training. EURASIP Journal on Information Security, 2012 (1), 1-22.

3. Arachchilage N. A. G., & Love S. (2014). Security awareness of computer users: A phishing threat avoidance perspective. Computers in Human Behavior, 38, 304-312.

https://www.metacompliance.com/lp/ultimate-guide-phishing.

Barry Schwartz. (2021). Google webspam report: Google Search Found 25 Billion Spammy Pages Each Day. Google Search Engine.

https://www.aol.com/

Tandale, K. D., & Pawar, S. N. (2020, October). Different types of phishing attacks and detection techniques: A review. In 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), pp. 295-299.

Alkhalil Z., Hewage C., Nawaf L., & Khan I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.

Ferreira R. P., Martiniano A., Napolitano D., Romero M., Gatto D. D. D. O., Farias E. B. P., & Sassi R. J. (2018). Artificial neural network for websites classification with phishing characteristics.

Haykin S. (2001) Redes Neurais-Princípios e Práticas. 2nd Edition, Bookman, Porto Alegre.

M. N. Badadhe, M. S. More, and N. V Puri, "An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms With Results," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). vol. 3, no. 5, pp. 1584–1589, 2014.

M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. -E. -. Ulfath and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1173-1179, doi: 10.1109/ICSSIT48917.2020.9214225.

Pandey A., Gill N., Sai Prasad Nadendla K., & Thaseen I. S. (2018, December). Identification of phishing attack in websites using random forest-svm hybrid model. In International conference on intelligent systems design and applications (pp. 120-128). Springer, Cham.

A. Yasin and A. Abuhasan, "An Intelligent Classification Model For Phishing Email Detection," Int. J. Netw. Secur. Its Appl., vol. 8, no. 4, 2016

Mhaske-Dhamdhere, V., & Vanjale, S. (2018). A Novel Approach for Phishing Emails Real Time Classification Using K-Means Algorithm. International Journal of Electrical & Computer Engineering (2088-8708), 8(6).

Tyagi, I., Shad, J., Sharma, S., Gaur, S., & Kaur, G. (2018, February). A novel machine learning approach to detect phishing websites. In 2018 5th International conference on signal processing and integrated networks (SPIN) (pp. 425-430). IEEE.

Deshpande, A., Pedamkar, O., Chaudhary, N., & Borde, S. (2021). Detection of phishing websites using Machine Learning. International Journal of Engineering Research & Technology (IJERT), 10(5).

Sönmez, Y., Tuncer, T., Gökal, H., & Avcı, E. (2018, March). Phishing web sites features classification based on extreme learning machine. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-5). IEEE.

A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," RTEICT 2017 – 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.

R. Rajesh and S. Prakash, "Extreme learning machines- a review and stateof-the-art," International Journal of Wisdom based Computing, vol. 1, no. 1, pp. 35-49, 2011.

Q. -Y. Zhu, A. K. Qin, P. N. Suganthan, G. -B. Huang, "Evolutionary Extreme Learning Machine," Pattern Recognition, vol. 38, pp. 1759-1763, 2005.