Feature Selection Technique for Improving Classification Performance in The Web-Phishing Detection Process

Anggit Ferdita Nugraha[1], Dwiky Alfian Tama[2], Dewi Anisa Istiqomah[3], Surya Tri Atmaja Ramadhani[4], Bayu Nadya Kusuma[5], Vikky Aprelia Windarni[6]
Faculty of Computer Science, University of Amikom Yogyakarta[1,2,3,4,5,6]
Jl. Padjajaran, Ring Road Utara, Condong Catur, Depok, Sleman, Yogyakarta[1,2,3,4,5,6]
Indonesia[1,2,3,4,5,6]
e-mail: anggitferdita@amikom.ac.id[1], dwiky.tama@students.amikom.ac.id[2], dewianisaist@amikom.ac.id[3], surya@amikom.ac.id[4], bayu.nadya@amikom.ac.id[5], vikkyaprelia@amikom.ac.id [6]

To cite this document:
Nugraha, A. F. ., Tama, D. A. ., Istiqomah, D. A. ., Ramadhani, S. T. A. ., Kusuma, B. N. ., & Windarni, V. A. . (2022). Feature Selection Technique for improving classification performance in the web-phishing detection process. Conference Series, 4(1), 25–31. https://doi.org/10.34306/conferenceseries.v4i1.667

Hash : ABCRcxeVygKM3n9LZbpfV33Aj4w6dJLrcEcl0tQunhAf6QUJWNPU6HQn2gDh5Myr

Abstract

Web phishing is a type of cybercrime that occasionally threatens the online activities of website visitors. Web phishing uses a phoney website page that closely mimics the legitimate Website in order to fool its target into providing crucial information. Web phishing attacks also continue to grow in popularity year after year. As a result, it is vital to design a web phishing detection system in order to reduce the number of victims and financial losses caused by web phishing attacks. The development of a web phishing detection system continues to this day, with machine learning being the most often used model. Unfortunately, the construction of a machine learning-based web phishing detection system frequently employs only a single classification step; however, the feature selection process enables an increase in the performance of the resultant classification. Thus, an experiment was conducted in this paper by using a feature selection procedure based on the Pearson correlation algorithm prior to doing machine learning modelling utilizing popular algorithms such as Naive Bayes, Decision Tree, and Random Forest. As a result, using a web phishing dataset from the UCI Machine Learning Repository, it was determined that the addition of the feature selection process based on the use of decision tree and random forest algorithms resulted in an increase in accuracy of up to 94.60 percent and 95.50 percent, respectively, and a slight decrease in accuracy of 0.4 percent when implemented in the Naive Bayes algorithm.

Keywords: Web-phishing, Feature Selection, Pearson correlation, Classification

p-ISSN : 2774-9576
e-ISSN : 2747-2981

1. Introduction

The Covid-19 epidemic alters a person's lifestyle and culture [1]. Work, study, and even shopping is increasingly being replaced by an internet society. This societal shift is inextricably linked to the Internet's function as a universal necessity [2]–[4]. With the Internet, one may readily obtain information and conduct numerous transactions. Behind the convenience, there are numerous cybercrime risks waiting to assault and harm internet users. Web phishing is an example of cybercrime that still poses a hazard to internet users, especially those who often deal online [5].

Web phishing is a cybercrime that attempts to deceive the target into divulging sensitive information. A false website page is created to look like the original website page, so the target does not realize he has been captured and his personal information such as login, password, or account number has been taken over and exploited by cybercriminals. This phishing web mainly targets transactional websites such as financial, e-commerce, airline and travel, and banking websites [6].

The number of web phishing websites is 611877 through mid-2021 and is expected to rise to 54% by early 2022 [7]. To reduce the number of web phishing victims, a system that can evaluate and detect web phishing attacks is required. Machine learning is still commonly utilized in phishing site detection systems. In web phishing research, algorithms like decision trees [8]–[10], Naive Bayes, and random forests [4], [8]–[10] are commonly utilized. Unfortunately, most machine learning research models only have one classification. As a result, the selected features are dominant and have a significant impact on the data classification process [11], [12]. So, in this study, we will use selection to see if it improves the performance of machine learning models, notably for web phishing detection. This experiment will use an online phishing dataset that can be downloaded for free from the UCI Machine Learning Repository.

2. Research Method

As indicated in Figure 1, this research was carried out in steps starting with preparation, modelling, data analysis, and model evaluation.

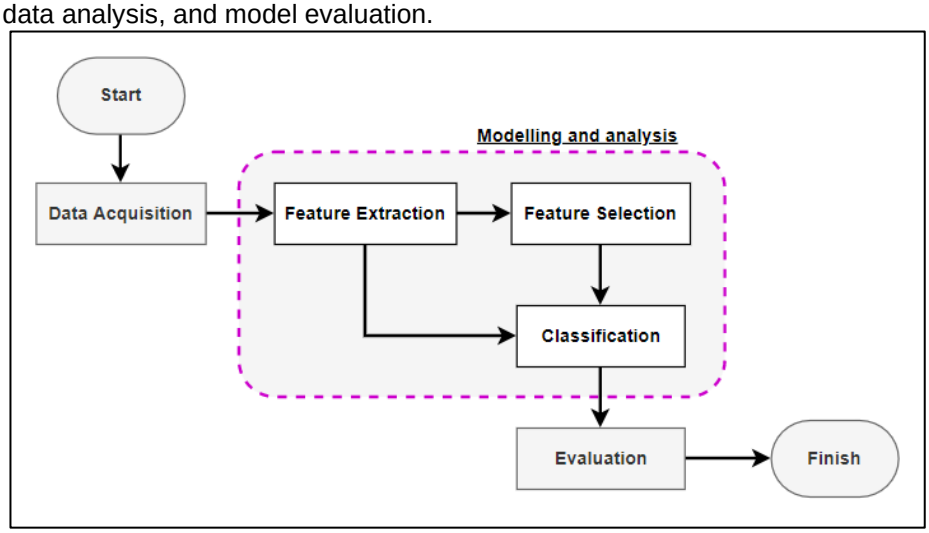

Figure 1. Research Flow.

Preparation involves downloading an online phishing dataset from the UCI Machine Learning Repository. The dataset contains 11055 data points with 30 features grouped into 4 categories: address bar-based features, abnormal-based features, HTML and JavaScript-based features, and domain-based features.

Table 1 provides details and explanations of each of these features.

Table 1. Web Phishing Dataset Features.

Section	Feature Name	Description
Address Bar based Features	Having IP Address	A website that is indicated as a phishing web uses an IP address in hexadecimal format.
	URL Length	If the character length of the URL address exceeds 54 characters, the Website is indicated as a phishing web.
	Shortening Service	Websites are indicated as phishing when using Short URLs such as "Tinyurl".
	Having at (@) symbol	If the Website contains the @ symbol in the URL, then the Website is indicated as phishing.
	Double Slash (//) redirecting	A legitimate website will place a double slash at the 6th character for HTTP and the 7th character for HTTPS.
	Prefix Suffix	A phishing web will use prefixes and suffixes, especially in the URL address.
	Having Sub Domain	Too many dots (.) in the URL address is one of the characteristics of a website is web phishing.
	SSL Final State	A valid website, of course, has an SSL certificate and uses HTTPS.
	Domain Registration Length	Web phishing tends to use the domain in a short time.
	Favicon	If the favicon is taken from an external domain, then the Website is a phishing website.
	Port	The Legitimate Website uses port 80 as a path for communication. In comparison, web phishing will open a port other than port 80 (HTTP).
	HTTPS Token	A phishing website does not have an authentication token like a legitimate website.
Abnormal based Features	Request URL	Media on the Legitimate Website are in the same URL and Domain.
	URL of Anchor	The <a> tag shows how many links are linked through the Website. The more links that are connected, the more the web is indicated as a phishing web.
	Links in Tags	Legitimate websites use <meta> tags for metadata, <script> tags for client-side scripting, and <link> for linking with other links.

	Server from Handler (SFH)	In web phishing, SFH is either blank or on another website.
	Submitting to Email	Phishing webs frequently use suspicious methods for sending data transmission links via email.
	Abnormal URL	Legitimate Websites are registered in the WHOIS database.
HTML and JavaScript-based Features	Redirect	Web phishing often does a lot of URL redirects.
	On MouseOver	Web phishing uses events to change the status bar.
	Right Click	Web phishing frequently disables right-click to prevent users from reading the source page.
	Popup Windows	Suspicious popup windows often appear on websites that are indicated as phishing.
	Iframe	If there are other websites that are hidden using Iframe, then the Website is indicated as a phishing website.
Domain Based Features	Age of Domain	Phishing websites have domain lifetimes of less than 6 months.
	DNS Record	Web phishing has DNS that is not recorded in the WHOIS database.
	Web Traffic	Traffic from phishing webs often looks suspicious.
	Page Rank	95% of phishing websites will not be found on search page rank.
	Google Index	Legitimate websites will be indexed by Google.
	Link Pointing to Page	A legitimate Website will display relationships with other websites.
	Statistical Report	In the category of top phishing IP or top phishing domain based on statistical reports from PhishTank or StopBadware.

The values applied for each feature are 1, 0, and -1. A website with a value of 1 belongs to the legitimate website category, whereas a feature with a value of 0 indicates that the Website is suspected as a web phishing website, and a feature with a value of -1 indicates that the Website is a phishing website.

The feature selection technique is carried out using Pearson correlation in the next stage, modelling and analysis. Pearson correlation is included in the category of feature selection filters because it performs a feature selection process based on the relationship between features and a data ranking process to determine the dominant feature. Pearson correlation transforms data into three different types of correlation coefficient values. A weak correlation is indicated by a value of 0, a strong and positive correlation by a value of 1, and a high but negative correlation by a value of -1.

The following equation is used to carry out the feature selection procedure using Pearson Correlation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where r is the value of correlation coefficient. x_i and y_i is the x-value and y-value at the i-th point, and then x dan y is the average of all values in the x and y. In addition to the number of features, the correlation coefficients for each feature are sorted. A correlation coefficient greater than or equal to 0.1 is considered as a dominant feature and used for processing in the next stage.

The following stage, data modelling and analysis, will represent all data on features that meet the chosen correlation coefficient threshold using popular machine learning methods such as Naive Bayes, Decision Tree, and Random Forest.

The outcomes of modelling using a well-known machine learning algorithm will then be analyzed to determine the improvement in classification performance caused by the feature selection process. The evaluation is conducted using a confusion matrix, as seen in Figure 2.

Actual Classification	**Prediction Classification**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Figure 2. Confusion matrix.

Figure 2 shows a confusion matrix for binary classification in which information about the experimental validity distribution between the actual data and the data produced from the prediction system based on the resulting model is recorded. Confusion matrices are classified into four types and are used as references in determining measurement metrics, namely:

- True Positive (TP): is a category in which the actual value is positive, and the system predicts that it will be positive.
- True Negative (TN): is a category in which both the actual and predicted values are negative.
- False Positive (FP): is a category in which the actual value is negative, but the system predicts it to be positive.
- False Negative (FN): is a category in which the actual value is positive, but the system predicts that it will be negative.

Using these 4 categories, the performance evaluation will be calculated based on the accuracy metric using the following equation:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3. Findings

The dataset studied in the research is the "Web Phishing Dataset" acquired from the UCI Machine Learning Repository. As the strategy mentioned previously, the research process will focus on the addition of a feature selection process prior to modelling utilizing popular machine learning methods. The study process is based on two (2) modelling scenarios. The first modelling is performed directly using a single classification process,

whereas the second modelling is performed by using the calculation of the correlation coefficient value based on the Pearson correlation as a feature selection procedure before the classification process. Pearson Correlation is used to determine the dominating characteristic, which is established by the correlation coefficient value above a predefined threshold.

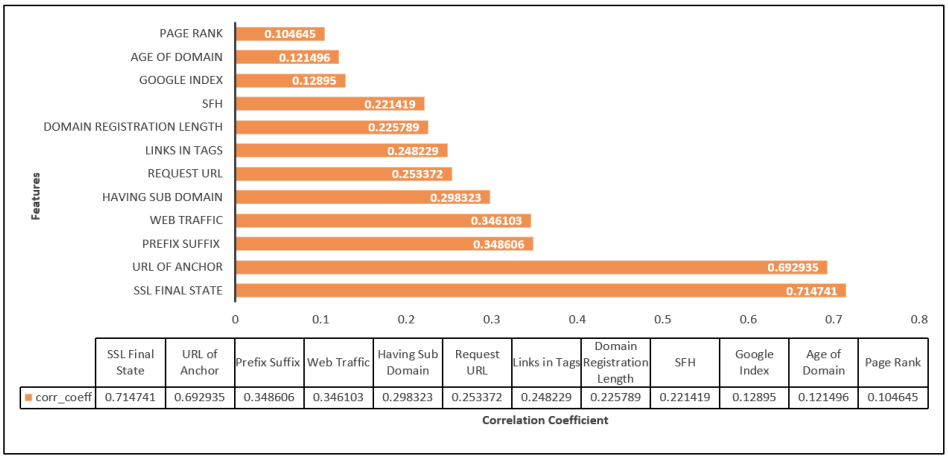

Figure 3. Correlation Coefficient value.

The correlation coefficient values calculated using the Pearson correlation equation are shown in Figure 3. There are 12 features with a correlation coefficient higher than the specified threshold of 0.1, with SSL Final State and URL of Anchor being 2 dominant features with a reasonably large gap when compared to features that have a value one level below like Prefix Suffix.

After that, when the classification performance is assessed in terms of the resulting accuracy value, the classification performance is as shown in Figure 4.

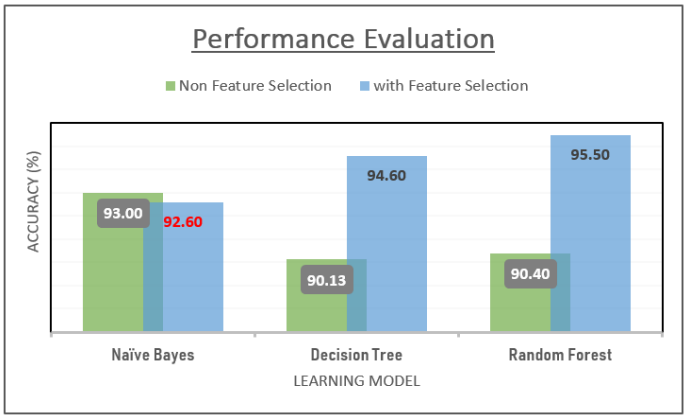

Figure 4. Comparison of accuracy values

Figure 4 shows the results of experiments conducted using three popular classification algorithms, that is Naive Bayes, Decision Tree, and Random Forest, and two different modelling scenarios, namely modelling without a feature selection process and modelling with a feature selection process based on the calculation of the correlation coefficient. As a result, the Naive Bayes method yields a classification accuracy of 93.00 % without the selection feature and 92.60 % with it. Using Decision Tree algorithm, the first model scenario achieves 90.13 % accuracy, while the second scenario, which includes a feature selection procedure, achieves up to 94.60 %. In experiments using the Random Forest algorithm, accuracy improved from 90.40 % in the first modelling scenario to 95.50 % when the feature selection process was added. Adding feature selection improves classification performance,

especially in Decision Tree and Random Forest algorithms. Unfortunately, the Naive Bayes algorithm's classification performance looks to have fallen slightly.

4. Conclusion

The purpose of this research was to determine whether adding a feature selection method prior to conducting machine learning modelling will improve the classification model's performance. Pearson Correlation was chosen as the feature selection approach, utilizing an online phishing dataset acquired from the UCI Machine Learning Repository. It was possible to extract up to 12 features out of 30 in the dataset based on the correlation coefficient value exceeding the stated threshold.

The performance of the feature selection process also improved, while in studies using the Naive Bayes algorithm, the classification performance actually declined, although only by 0.4 percent. While compared to the reduction in performance, the enhanced accuracy value has a higher proportion of values between 4 and 5%, with the best accuracy value of 95.5 % achieved when applying Random Forest as the classification algorithm.

References

[1] S. Mulyaningsih, L. Amalia, and H. Hernawan, “Edukasi Adaptasi Kebiasaan Baru Pada Masa Pandemi Covid-19,” *J. PEKEMAS*, vol. 3, pp. 5–8, 2020.

[2] A. Kulkani and L. L. Brown, “Phishing websites detection using machine learning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 8–13, 2019.

[3] R. Kiruthiga and D. Akila, “Phishing websites detection using machine learning,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, pp. 111–114, 2019.

[4] S. Parekh, D. Parikh, S. Kotak, and S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 949–952.

[5] G. Harinahalli Lokesh and G. BoreGowda, “Phishing website detection based on effective machine learning approach,” *J. Cyber Secur. Technol.*, vol. 5, no. 1, pp. 1–14, 2021.

[6] M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” *6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding*, vol. 2018-Janua, pp. 1–5, 2018.

[7] J. Johnson, “Phishing - statistics & facts,” 2021. [Online]. Available: https://www.statista.com/topics/8385/phishing/#dossierKeyfigures. [Accessed: 19-Jan-2022].

[8] X. Yang, L. Yan, B. Yang, and Y. Li, “Phishing Website Detection Using C4.5 Decision Tree,” *DEStech Trans. Comput. Sci. Eng.*, no. itme, pp. 119–124, 2017.

[9] L. Machado and J. Gadge, “Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017, pp. 1–5.

[10] D. R. Patil and J. B. Patil, “Malicious URLs detection using decision tree classifiers and majority voting technique,” *Cybern. Inf. Technol.*, vol. 18, no. 1, pp. 11–29, 2018.

[11] A. F. Nugraha and L. Rahman, “Meta-algorithms for improving classification performance in the web-phishing detection process,” *2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019*, vol. 6, pp. 271–275, 2019.

[12] Y. Pristyanto, I. Pratama, and A. F. Nugraha, “Data level approach for imbalanced class handling on educational data mining multiclass classification,” in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 310–314.