Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12958 p-ISSN : 2541-044X

Performance of Various Naïve Bayes Using GridSearch Approach In Phishing Email Dataset

Rizki Rahman[1)*], Ferian Fauzi Abdulloh[2)]
[1,2)]Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[1)]riskirahman1116@students.amikom.ac.id, [2)]ferian.fauzi@amikom.ac.id

Submitted : Sep 1, 2023 | **Accepted** : Sep 16, 2023 | **Published** : Oct 1, 2023

Abstract: The background is the increasing cybersecurity threats in the form of phishing attacks that can be detrimental to individuals and organizations. The purpose of this research is to compare the performance of four Naive Bayes variants in classifying phishing emails with a method that involves a data pre-processing stage, phishing emails are collected, cleaned, and converted into appropriate numerical features. Next, the GridSearch approach was used to find the best parameters. This research objective is to understand how each Naive Bayes variant works on phishing email datasets. This phishing detection task is based on the following performance evaluation criteria such as accuracy, precision, recall, and F1-score. In this study, Bernoulli got the best accuracy of 97.34% but when the results obtained a hyperparameter, the results showed an increase with the most optimal results and the best performance is Bernoulli 97.38%. The research results are to provide an in-depth insight into the effectiveness of each variant of Naive Bayes in dealing with phishing email datasets and researchers in selecting the most suitable Naive Bayes variant for phishing detection tasks. In addition, the applied GridSearch method can guide how to find the best parameters for Naive Bayes models in other contexts. In summary, this study focuses on analyzing the performance of four variants of Naive Bayes Gaussian, Multinomial, Complement, and Bernoulli with the best algorithms Bernoulli 97.38%.

Keywords: Bernoulli, Gaussian, GridSearch, Naïve Bayes, Phishing Email

INTRODUCTION

In an increasingly advanced digital era, cybersecurity threats are becoming increasingly complex and troubling. One of the most common and harmful forms of attack is a phishing attack, especially in the form of a phishing email. These attacks make use of psychological and technical manipulation techniques to lure users into disclosing personal information, financial data, and other sensitive information. Phishing emails are becoming the most common tool used by cybercriminals due to their easy nature to spread and the ability to fool even the most vigilant users (Hadi Ramadhan & Kumalasari Nurnawati, 2022).

Phishing email attacks create serious challenges for individuals, companies and other institutions. Various attempts have been made to protect users from these attacks, including the use of advanced technology and security awareness education. Nonetheless, email phishing attacks continue to evolve with increasingly sophisticated tactics that are difficult to detect. Therefore, a more advanced approach is needed to deal with this threat (Hadi Ramadhan & Kumalasari Nurnawati, 2022).

One solution that has been widely used in phishing attack detection is the application of data analysis and machine learning techniques. Classification methods, such as Naive Bayes, have proven effective

*name of corresponding author

in identifying patterns that might signal the presence of phishing emails. Naive Bayes variants, such as Gaussian, Multinomial, Complement, and Bernoulli, offer a unique approach to dealing with categorical data in the context of phishing analysis. Researchers classify email spam or non-spam using logistic regression. Email data that has been obtained is preprocessed and then logistic regression modeling is performed. Logistic regression performance was obtained by 95% and Naïve Bayes was obtained by 93% (Suprihati, 2021). So, I want to make a difference in the algorithm and the difference in the phishing email dataset from previous research

However, the effectiveness of each variant of Naive Bayes in dealing with phishing email datasets still needs to be studied further. Therefore, this study aims to analyze the performance of the four variants of Naive Bayes with the GridSearch approach on phishing email datasets. GridSearch was picked because it helps automate the process of finding the best parameters by systematically testing different combinations of parameters. This reduces the need to manually try parameters, which can be time and resource consuming, and can result in a model with optimal performance. GridSearch is particularly useful for Naive Bayes and is related in that it can find the most common parameter in Naive Bayes, alpha, which controls the Laplace smoothing used to solve the zero probability problem. It is hoped that the results of this study will provide deeper insight into the advantages and limitations of each variant in classifying phishing emails. The stages of this method itself start from data collection, data preprocessing, TF-IDF, model selection, Gridsearch approach, and evaluation. The targeted output of this research is a better understanding of how each variant of Naive Bayes works on phishing email datasets. It is hoped that it will be revealed which variant is most suitable for this phishing detection task based on performance evaluation criteria such as accuracy, precision, recall, and F1-score.

With a better understanding of the performance of Naive Bayes variants in the context of phishing email detection, it is hoped that more effective and adaptive cybersecurity measures can be implemented. This research can contribute to the development of early detection techniques against email phishing attacks, which in turn will help protect users' personal and sensitive information as well as overall organizational security and provide in-depth insights into the effectiveness of each variant of Naive Bayes against phishing email datasets. The research results can provide practical guidance for cybersecurity professionals and researchers in selecting the most suitable variant of Naive Bayes for the phishing detection task. In addition, the applied GridSearch method can guide how to find the beast parameters for the Naive Bayes model in other contexts.

LITERATURE REVIEW

This study makes comparisons between naïve Bayes variants and approaches through GridSearch validation as hyperparameters from phishing email data to find the best parameters of the naïve Bayes model. In this modeling used classification which is the process of finding a class model to be categorized (Suprihati, 2021). Naïve bayes itself is a method of determining probabilities and predicting opportunities using data. The label given is the target to be addressed (Momole, 2022).

Research by (Darmawan & Fauzan Dianta, 2023), Researchers use the algorithm, namely SVM. In this study, researchers predicted heart disease. With the SVM algorithm, researchers got 83% accuracy. While the results from GridSearch get an increase of 86%. Research by (Momole, 2022), researchers compared algorithms, namely naïve Bayes and random forest to get the best performance. In this study, researchers only made comparisons without optimization and looked for the best parameters. The results of this study got a naïve Bayes algorithm of 99.22%. Research by (Kurniadi et al., 2022), researchers conducted research on predicting student data using the naïve Bayes algorithm. From this research, 3 tests were carried out, namely when using the naïve Bayes algorithm, using the forward selection feature, then SMOTE. Maximum results are obtained from the third model, naïve bayes using forward selection and SMOTE features with a performance accuracy of 87.13%. Research by(Putri & Wijayanto, 2022), researchers classify phishing website use the Naïve Bayes algorithm, Decision tree, Random forest and SVM with the best Random forest algorithm of 90.77%. Research by (Subarkah & Ikhsan, 2021), researchers detection phishing website use the CART algorithm and gets an accuracy 95.28%. Research by (Suprihati, 2021) ,researchers classify email spam or non-spam using logistic regression. Email data that has been obtained is preprocessed and then logistic regression modeling is performed. Logistic

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12958 p-ISSN : 2541-044X

regression performance was obtained by 95%. Research by (Agus Fatkhurohman, Eli Pujastuti, 2019), researchers use naïve bayes with cross validation and the final accuracy naïve bayes is 92.98%.

METHOD

This research aims to analyze the performance of Naïve Bayes regarding Email Phishing. This Phishing Email is obtained from open source Kaggle with name Phishing Email Detection. In this dataset aim to trick recipients into divulging sensitive information or performing harmful actions. The dataset specifies the email text body the type of emails which can be used to detect phishing emails. The data contains two features namely Email Text and Email Type. The Email Text specifies the email body and the Email Type specifies the type of email whether it is Phishing or Safe. The results of this study can be used as a comparison of naïve bayes variants with the GridSearch approach is expected to be a significant performance improvement. To carry out this research, I used google colab with the research flow in the figure below.

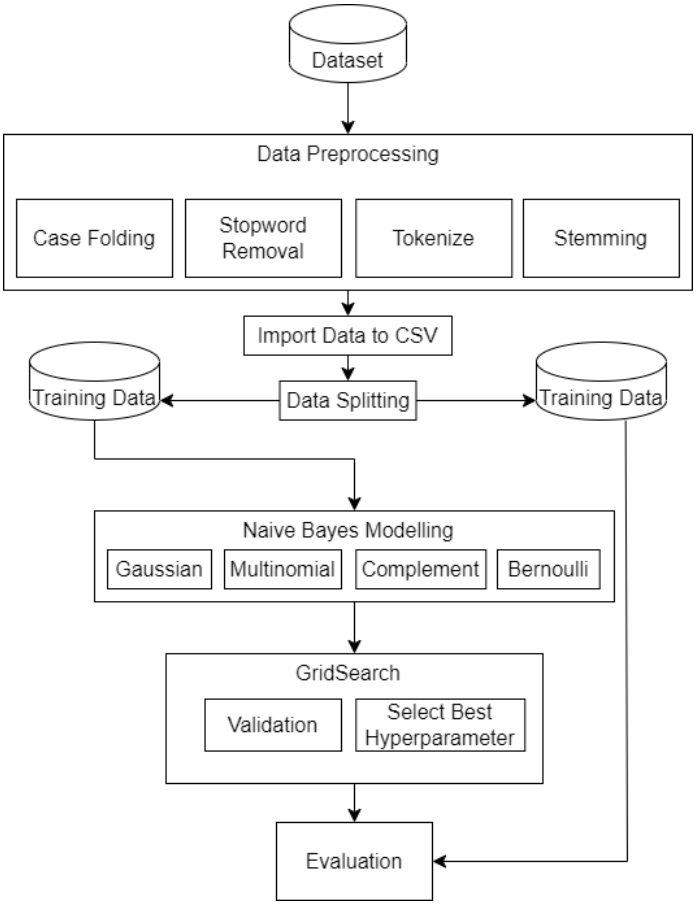

Fig 1. Framework of Research

Dataset

Researchers collect data from open source datasets on kaggle with the title phishing email. Researchers get data related to safe email text data and phishing emails in the form of CSV file datasets containing email lists. This data attribute contains email text which is the email text and email type which contains the type of phishing or secure email. The participant data registered in the excel file amounted to 171,004 data contains text sent via email with a collection of words, letters, and sentences. There are only 2 variables in this data, Email Text and Email Type. The amount of data from these variables does not contain empty data and has the same amount. As much text data as there is has an average email text length and different email type frequencies. so this data needs to be processed so that it can be analyzed.

Case Folding

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12958 p-ISSN : 2541-044X

Case folding here is used in text processing to remove punctuation, double spaces, and numbers. Then each word that has been processed is converted into lowercase letters. In addition to converting characters to lowercase, the case folding procedure uses the regular expression library to remove numbers, punctuation marks, and emoticons. After removing numbers, punctuation marks, and emojis, researchers removed duplicate text.

Stopword Removal

Stopwords are used in a language but generally do not have significant meaning or contribute to the overall understanding of the text. Stopwords usually include words like "the", "and", "is", "in", "of", "it", "to", "on", "as", "with", and many more, depending on the language. These words serve a grammatical purpose such as connecting words, indicating tenses, or building sentence structures. However, they have no specific semantic content and can often be safely omitted without affecting the core meaning of the text. This technique is used to reduce noise and improve processing efficiency.

Tokenize

The tokenizing process functions to convert a sentence into every word that makes up the sentence the sentence. For example in the sentence "I will go on vacation" it will become "I", "will", "go", "on vacation" by using the tokenizing method.

Stemming

Stemming is the process of processing text processing that serves to cut affixes from each word and make
each word into a base word.

Split Data

Preprocessed data is divided into 80% training data and 20% testing data. After splitting the data, TF-IDF is performed. The TF-IDF approach presents text with a vector space in which each feature in the text corresponds to one word. TF (Term Frequency) will calculate the frequency of occurrence of a word and compared to the total number of words in the document (Al-talib & Hassan, 2013). Formula for TF-IDF is:

$$W_{dt} = tf_d * idf_t = tf_d * \log\left(\frac{N}{df_t}\right) \quad (1)$$

Gaussian Naïve Bayes

The Gaussian distribution, also known as the normal distribution, is the most commonly used probability distribution in statistical analysis. It is often used to describe data that is symmetrically distributed around its mean. The Gaussian distribution has a bell-shaped curve and can be fully described by two parameters, the mean and standard deviation (Karunia et al., 2017).

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{C,i}^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2)$$

Multinomial Naïve Bayes

The Multinomial distribution is used to describe the probability distribution of an experiment that has more than two possible outcomes. It is commonly used in categorical data analysis. The Multinomial distribution is a generalization of the binomial distribution that applies to more than two categories (Karunia et al., 2017).

$$P(X - x) = \frac{n!}{x_1! * x_2! \ldots x_k!} * P_1{}^{x_1} * P_2{}^{x_2} \ldots P_k{}^{x_k} \quad (3)$$

Complement Naïve Bayes

Complement Naïve Bayes is a variation of the Naïve Bayes classification algorithm specifically designed to address imbalances in classification datasets. Complement Naïve Bayes uses the opposite approach to regular Naïve Bayes. It considers the proportion of the frequency of each attribute in a class that does not correspond to that class (Karunia et al., 2017).

$$P(C|x) = \frac{P(x|C) * P(C)}{\sum_{c'} P(x|c') * P(c')} \quad (4)$$

Bernoulli Naïve Bayes

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12958 p-ISSN : 2541-044X

The Bernoulli distribution is used to describe an experiment that has two possible outcomes, such as success or failure, yes or no. The Bernoulli distribution represents the probability distribution of success in a single Bernoulli experiment (Karunia et al., 2017).

$$P(X = x) = p^x * (1 - p)^{1-x} \quad (5)$$

GridSearch & Evaluation

GridSearch is one of the methods used in selecting the best parameters for a machine learning model or algorithm. The goal is to find a combination of parameters that produces optimal model performance by comparing the results of various combinations. Parameter combinations are entered in the evaluation. The combination selected is the one that gives the highest value to the evaluation metric.

RESULT

The data to be used amounted to 171,004 data. After that, preprocessing is carried out so that clean from noise or words that are not necessary. This process consists of several parts, namely case folding, tokenizing, filtering, and stemming. Clean data is split with 80% training data and 20% testing data. Furthermore, we obtained email type data with a new number of 17,853. Raw data in phishing email before preprocessing :

Table 1 Raw Data Phishing Email

Email Text	Email Type
re : 6 . 1100 , disc : uniformitarianism , re ...	0
the other side of * galicismos * * galicismo *...	0
re : equistar deal tickets are you still avail...	0
\nHello I am your hot lil horny toy.\n I am...	1
software at incredibly low prices (86 % lower...	1
…..	….

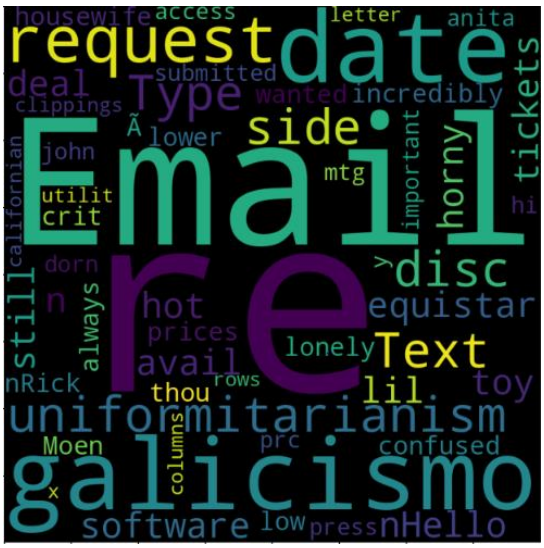

Fig 2 Frequently occurring data before preprocessing

Data that has been collected will be done in case folding, stopword removal, tokenizing, and stemming. Data after stemming is data that will be used in the model.

Table 2 The result processing of Case Folding, Stopword Removal, Tokenizing, Stemming

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12958 p-ISSN : 2541-044X

Email Text	Email Type	Case Folding	Stopword	Tokenizing	Stemming
re : 6 . 1100 , disc : uniformitarianism , re ...	0	re disc uniformitarianism re sex la...	disc uniformitarianism sex lang dick hudson ob...	[disc, uniformitarianism, sex, lang, dick, hud...	disc uniformitarian sex lang dick hudson obser...
the other side of * galicismos * * galicismo *...	0	the other side of galicismos galicismo is ...	side galicismos galicismo spanish term names i...	[side, galicismos, galicismo, spanish, term, n...	side galicismo galicismo spanish term name imp...
…	…	…	…	…	…

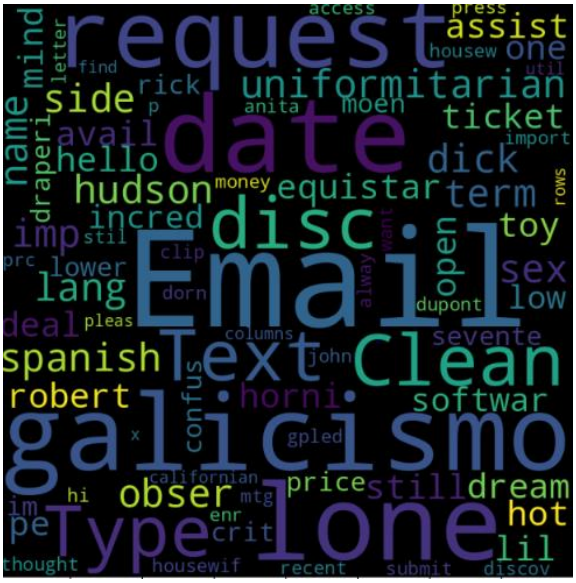

Fig 3 Frequently occurring data after preprocessing

After that, the data stemming is imported into a new CSV, and Naive Bayes Classifier modeling is carried out. At this stage, researchers used 80% training data and 20% testing data. Metrics used to evaluate the model include accuracy, precision, recall, F1-Score, and balance accuracy. Balance accuracy is used because of the imbalance in the amount of email types.

	accuracy	f1_score	precision	recall	balanced_accuracy
Bernoulli	0.972626	0.962820	0.949580	0.976434	0.973445
Complement	0.969775	0.958431	0.956930	0.959937	0.967659
Multinomial	0.964072	0.948905	0.980721	0.919089	0.954397
Gaussian	0.886228	0.826446	0.925926	0.746269	0.856124

Fig 4 The result of metrics before hyperparameter tuning

From these results, researchers try to combine parameters so that the results obtained from the four models get maximum results. Here researchers use GridSearch with **alpha parameters: [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0], fit_prior: [True, False], force_alpha:[True, False], and var_smoothing: [1e-9, 1e-8, 1e-7, 1e-6, 1e-5, 12-4, 1e-3, 1e-2, 1e-1].** The results of the combination search will be incorporated into the model and re-evaluated.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12958
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Bernoulli
Best hyperparameter: {alpha : 0.5, fit_prior : True, force_alpha : True}
Best mean cross-validated score : 0.9735510211966816

Complement
Best hyperparameter: {alpha : 0.1, fit_prior : True, force_alpha : True}
Best mean cross-validated score : 0.973978727060314

Multinomial
Best hyperparameter: {alpha : 0.1, fit_prior : False, force_alpha : True}
Best mean cross-validated score : 0.973978727060314

Gaussian
Best hyperparameter: {var_smoothing : 0.01}
Best mean cross-validated score : 0.9498813425073438

After completing the hyperparameter stage, the researcher implemented the model again but used the parameters that had been obtained previously. The data built produces phishing email data and safe email data. Figure 6 illustrates the graph of the amount of email types

	accuracy	f1_score	precision	recall	balanced_accuracy
Bernoulli	0.972911	0.963221	0.949618	0.977219	0.973838
Multinomial	0.970060	0.958936	0.954829	0.963079	0.968558
Complement	0.970060	0.958936	0.954829	0.963079	0.968558
Gaussian	0.949244	0.928341	0.952106	0.905734	0.939886

Fig 5 The result of metrics after hyperparameter tuning

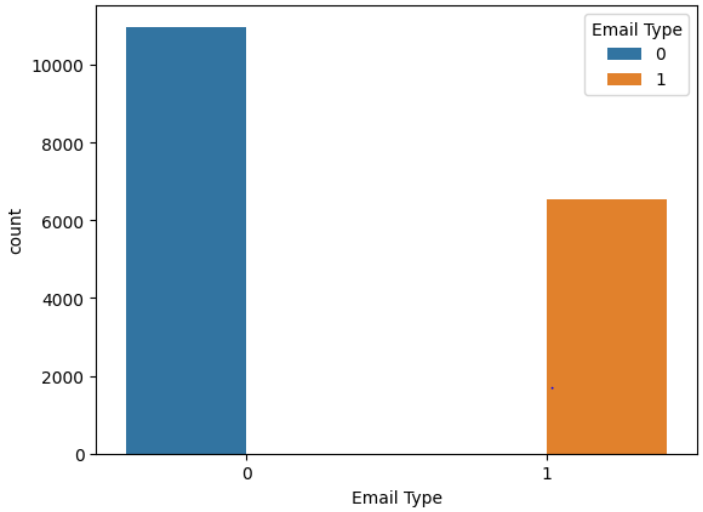

Fig 6 Dataset Visualization

Label with 0 indicating safe email type and 1 indicating phishing email type.

DISCUSSIONS

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12958 p-ISSN : 2541-044X

From the results of the research that has been done, at this stage an evaluation is carried out so that it is known to increase the value of each Naïve Bayes model before using GridSearch and after using GridSearch by searching for a combination of parameters from each model. The following is a comparison of the balance accuracy value of each Naïve Bayes model before using GridSearch and Naïve Bayes after using GridSearch.

Table 3 Comparison of Performance Improvements Results

	Gaussian	Multinomial	Complement	Bernoulli
Before GridSearch	85.61%	95.43%	96.76%	97.34%
After GridSearch	93.99%	96.86%	96.86%	97.38%
Increased	**+8.38%**	**+1.43%**	**+0.1%**	**+0.04%**

In Table 3, it can be seen that each model has improved performance after hyperparameterization. So it is proven that model parameter optimization can improve the performance of each Naïve Bayes model in email phishing classification. This research is almost similar to previous research but uses spam email datasets. Previous researchers used logistic regression and additional methods, namely naive bayes. In this study, logistic regression got the best performance of 95% and naive bayes 93% (Ferin Reviantika, Yufis Azhar, Gita Indah Marthasari, 2021).

The limitations of this research are only limited to classifying Phishing emails and using four variants of naive bayes.

CONCLUSION

Researchers concluded that the results of the email phishing classification process using the Naïve Bayes Classifier algorithm can help classify data that includes phishing or safe data. The best performance results obtained from each model in the study are Gaussian 93.99%, Multinomial 96.86%, Complement 96.86%, and Bernoulli 97.38%. The classification results show an increase in performance when using GridSearch, some make better improvements such as Gaussian. The algorithms that improved and made the performance the same were Multinomial and Complement. Before hyperparameterization, the Bernoulli algorithm had the best performance. After the hyperparameter is done, Bernoulli's algorithm still gets the best performance but with an increase of 97.38%. It can be concluded from the comparison of these algorithms that Bernoulli is very effective in classifying Phishing Emails and the GridSearch approach is very useful in improving the Naive Bayes model. In previous research, research was conducted on email spam using logistic regression and naive bayes. It is shown that Logistic Regression is superior to Naive Bayes, which is 95%. So this research uses the Phishing Email dataset which is similar to the previous researcher's data and gets better performance than the previous researcher by 97.38% of the Bernoulli Naive Bayes model.

REFERENCES

Afdhaluzzikri, A., Mawengkang, H., & Sitompul, O. S. (2022). Perfomance analysis of Naive Bayes method with data weighting. *SinkrOn*, *7*(3), 817–821. https://doi.org/10.33395/sinkron.v7i3.11516

Al-talib, G. A., & Hassan, H. S. (2013). A Study on Analysis of SMS Classification Using TF-IDF weighting. *International Journal of Computer Networks and Communications Security*, *1*(2013), 189–194. https://doi.org/10.47277/ijcncs/1(5)3

Amin Muftiadi. (2022). Studi kasus keamanan jaringan komputer: analisis ancaman phisingterhadap layanan online banking. *Hexatech: Jurnal Ilmiah Teknik*, *1*(2), 60–65.

Bustomi, Y., Nugraha, A., Juliane, C., & Rahayu, S. (2023). Data Mining Selection of Prospective Government Employees with Employment Agreements using Naive Bayes Classifier. *Sinkron*, *8*(1), 1–8. https://doi.org/10.33395/sinkron.v8i1.11968

Christanto, B., & Setiabudi, D. H. (2020). Penerapan Random Forest dalam Email Filtering untuk Mendeteksi Spam. *Jurnal Infra*, *8*(2).

*name of corresponding author

Darmawan, Z. M. E., & Fauzan Dianta, A. (2023). Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM. *Online) Teknologi: Jurnal Ilmiah Sistem Informasi*, *13*(1), 8–15. https://doi.org/10.26594/teknologi.v13i1.3098Tersediaonlinediwww.journal.unipdu.ac.idHalamanjurnaldiwww.journal.unipdu.ac.id/index.php/teknologi

Fatkhurohman, A., & Pujastuti, E. (2019). Penerapan Algoritma Naïve Bayes Classifier Untuk Meningkatkan Keamanan Data Dari Website Phising. *Respati*, *14*(1), 115–124. https://doi.org/10.35842/jtir.v14i1.279

Hadi Ramadhan, I., & Kumalasari Nurnawati, E. (2022). Analisis Ancaman Phishing Dalam Layanan E-Commerce. *Prosiding Snast*, *November*, E31-41. https://doi.org/10.34151/prosidingsnast.v8i1.4169

Karunia, S. A., Saptono, R., & Anggrainingsih, R. (2017). Online News Classification Using Naive Bayes Classifier with Mutual Information for Feature Selection. *Jurnal Ilmiah Teknologi Dan Informasi*, *6*(1), 11–15. https://jurnal.uns.ac.id/itsmart/article/view/11114

Kurniadi, D., Nuraeni, F., & Lestari, S. M. (2022). Implementasi Algoritma Naïve Bayes Menggunakan Feature Forward Selection dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana. *Jurnal Sistem Cerdas*, *5*(2), 63–82. https://doi.org/10.37396/jsc.v5i2.215

Lubis, A. I., & Chandra, R. (2023). Forward Selection Attribute Reduction Technique for Optimizing Naïve Bayes Performance in Sperm Fertility Prediction. *Sinkron*, *8*(1), 275–285. https://doi.org/10.33395/sinkron.v8i1.11967

Momole, G. M. (2022). Perbandingan Naïve Bayes dan Random Forest Dalam Klasifikasi Bahasa Daerah. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(2), 855–863. https://doi.org/10.35957/jatisi.v9i2.1857

Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika : Jurnal Sistem Komputer*, *11*(1), 59–66. https://doi.org/10.34010/komputika.v11i1.4350

Rahmadani, P. S., Tampubolon, F. C., Jannah, A. N., Hutabarat, N. L. H., & Simarmata, A. M. (2022). Tiktok Social Media Sentiment Analysis Using the Nave Bayes Classifier Algorithm. *SinkrOn*, *7*(3), 995–999. https://doi.org/10.33395/sinkron.v7i3.11579

Subarkah, P., & Ikhsan, A. N. (2021). Identifikasi Website Phishing Menggunakan Algoritma Classification And Regression Trees (CART). *Jurnal Ilmiah Informatika*, *6*(2), 127–136. https://doi.org/10.35316/jimi.v6i2.1342

Suprihati, F. R. (2021). Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression. *Jurnal Sistem Cerdas*, *4*(3), 155–160. https://doi.org/10.37396/jsc.v4i3.166

Tangkelayuk, A. (2022). The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(2), 1109–1119. https://doi.org/10.35957/jatisi.v9i2.2048

Tanjung, J. P., Tampubolon, F. C., Panggabean, A. W., & Nandrawan, M. A. A. (2023). Customer Classification Using Naive Bayes Classifier With Genetic Algorithm Feature Selection. *Sinkron*, *8*(1), 584–589. https://doi.org/10.33395/sinkron.v8i1.12182

*name of corresponding author