CENDERAWASIH Journal of Statistics and Data Science
Volume 1 Nomor 2 Juni 2023

Perbandingan Metode Klasifikasi *Decission Tree, Naive Bayes, K-Nearest-Neighbor,* dan *Logistic Regression* pada Dataset *Phishing*

WICKLY GUSTHVI[1], AFRIONALDI A. ROZA[2], CAECILIA BINTANG GIRIK ALLO[3]

[1)]Program Studi Statistika Fakultas MIPA Universitas Cenderawasih, Indonesia
[2)] Program Studi Statistika Fakultas MIPA Universitas Cenderawasih, Indonesia
[3)] Program Studi Statistika Fakultas MIPA Universitas Cenderawasih, Indonesia
e-mail: wickly23@gmail.com[1], afrionaldiroza19@gmail.com[2]

ABSTRAK

Dalam menentukan algoritma yang digunakan pada *machine learning*, diperlukan pengujian kinerja terhadap metode klasifikasi yang digunakan. Pada artikel ini akan dibandingkan kinerja dari metode klasifikasi *Decission Tree, Naive Bayes, K-Nearest-Neighbor*, dan *Logistic Regression* pada dataset *phising* sehingga dapat memberikan gambaran tentang metode yang relatif optimal untuk dipilih sebagai tahap awal dalam memilih algoritma yang akan diterapkan. Metode yang digunakan untuk mengukur kinerja empat model pada penelitian ini adalah nilai *accuracy*, kurva ROC dan nilai AUC. Semakin tinggi nilai *accuracy* dan nilai AUC maka semakin baik suatu model. Nilai *accuracy* dan nilai AUC yang dihasilkan pada model *logistic regression* paling tinggi dibandingkan tiga model lainnya.

Kata Kunci: Decission Tree, Naive Bayes, K-Nearest-Neighbor, Logistic Regression, Accuracy, ROC, AUC

1. PENDAHULUAN

Teknologi *machine learning* (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan *data mining* sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah. Dalam hal ini *machine learning* memiliki kemampuan untuk memperoleh data yang ada dengan perintah sendiri. ML juga dapat mempelajari data yang ada dan data yang ia peroleh sehingga bisa melakukan tugas tertentu. Tugas yang dapat dilakukan oleh ML pun sangat beragam, tergantung dari apa yang ia pelajari. Istilah machine learning pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar-dasar *machine learning* dan konsepnya. Sejak saat itu ML banyak yang mengembangkan.

Pada *machine learning* dikenal algoritma *machine learning*. Algoritma *machine learning* adalah metode dimana sistem *artificial intelligence* mengerjakan tugasnya secara otomatis. Umumnya algoritma *machine learning* ini digunakan untuk memprediksi nilai output dari input yang diberikan. Algoritma machine learning sendiri dibagi menjadi dua, yaitu supervised dan *unsupervised learning*. *Supervised learning* membutuhkan data input dan data output yang diinginkan dan digunakan untuk membuat pelabelan, sedangkan algoritma *unsupervised learning* bekerja dengan data yang tidak diklasifikasikan atau tidak diberi label. Contoh algoritma *unsupervised learning* adalah pengelompokan atau *clustering data* yang tidak difilter berdasarkan persamaan dan perbedaan. Sedangkan algoritma *supervised learning*, yaitu algoritma klasifikasi.

Terkadang sulit memutuskan algoritma *machine learning* mana yang paling baik untuk klasifikasi diantara banyaknya pilihan dan jenis algoritma klasifikasi yang ada. Namun, ada algoritma klasifikasi *machine learning* yang paling baik digunakan dalam masalah atau situasi tertentu. Algoritma klasifikasi ini digunakan untuk klasifikasi teks, analisis sentimen, deteksi spam, deteksi penipuan, segmentasi pelanggan, dan klasifikasi gambar. Pilihan algoritma yang sesuai bergantung pada kumpulan data dan tujuan yang akan dicapai. Sebagaimana tujuan dari penelitian ini untuk mengetahui kelas model klasifikasi yang lebih baik dan relatif optimal pada dataset phishing. Ada banyak metode klasifikasi yang telah dikembangkan namun kinerjanya selalu berbeda dari satu masalah ke masalah lain. Perlu dilakukan

suatu upaya untuk memilih model klasifikasi yang relatif optimal, salah satunya dengan membandingkan kinerja beberapa algoritma dari beberapa kelas model berbeda menggunakan dataset yang ada sehingga dapat memberikan gambaran tentang kelas model mana yang relatif optimal untuk dipilih sebagai tahap awal dalam memilih algoritma yang akan diterapkan. Pada penelitian ini dibandingkan kinerja metode klasifikasi dari model *Decission Tree, Naive Bayes, K-Nearest-Neighbor,* dan *Logistic Regression*.

2. METODE PENELITIAN

Secara sistematis penelitian ini dibagi ke dalam beberapa tahapan penelitian yaitu, yakni pengumpulan data, pemodelan, dan evaluasi.

2.1 Sumber Data

Data yang digunakan yaitu dataset *phishing* yang di ambil dari website kaggle. *Phising* merupakan upaya untuk mendapatkan informasi data seseorang dengan Teknik pengelabuan. Data yang menjadi sasaran *phising* adalah data pribadi (nama, usia, alamat), data akun (*username* dan *password*), dan data finansial (informasi karu kredit, rekening). Istilah resmi *phising* adalah *phishing* , yang berasal dari kata *fishing* yaitu memancing. Kegiatan *phising* memang bertujuan memancing orang untuk memberikan informasi pribadi secara sukarela tanpa disadari, padahal informasi yang dibagikan tersebut akan digunakan untuk tujuan kejahatan. Proses selanjutnya, peneliti mengolah seluruh data dari hasil yang sudah dikumpulkan dengan melakukan *cleansing* guna mengatasi masalah yang biasa ditemukan seperti nilai yang tidak ada, anomaly, data tidak sesuai atau adanya redundansi atau pengulangan data. Setelah itu, data dikelompokkan lalu dipilah sesuai dengan jenis dan fungsinya untuk di distribusikan ke dalam data *testing* dan *training* dengan perbandingan 70:30, lalu diterapkan didalam model-model klasifikasinya. Dari hasil pengolahan data, didapat sebesar 11430 *rows* dan 87 *columns.*

2.2 Model klasifikasi

Model klasifikasi yang digunakan dalam penelitian ini terdiri dari 4 model, yaitu *Decission Tree, Naive Bayes, K-Nearest-Neighbor,* dan *Logistic Regression.*

2.2.1 *Decission Tree*

Model *Decision Tree* dibentuk menyerupai struktur *flowchart*, dimana setiap simpul yang bukan simpul daun merupakan atribut pengujian, setiap cabang mewakili output dari pengujian, dan setiap simpul daun (*terminal node*) menentukan label *class*. Simpul paling atas dari sebuah pohon adalah node akar (Han et al., 2011).

Decision Tree membedakan antara suatu kelas dengan kelas lainnya berdasarkan tingkat kemurnian kelas (*impurity*) pada suatu simpul. Alat ukur kemurnian kelas (*impurity*) yang umum digunakan adalah GINI *Index, Entropy, Misclassification measure, Chi-square measure, G-square measure.* Beberapa algoritma yang termasuk kategori *Decission Tree* antara lain *ID3, C4.5, C5.0, CART, CHAID*, dan lain-lain (Gorunescu, 2011).

2.2.2 *Naive Bayes*

Algoritma *Naïve Bayes* mengasumsikan bahwa semua atribut adalah variabel independen. Asumsi ini disebut *class-conditional independence* sehingga komputasi algoritma ini menjadi sangat sederhana. Inilah yang menyebabkan model ini dinamakan dengan "*naïve"*. Teorema bayes dapat ditulis menggunakan persamaan 1 (Hadiwandra, 2019):

$$P(A|B) = \frac{P(B|A)\times P(A)}{P(B)} \quad \ldots(1)$$

Dimana :

$P(A|B)$ = Probabilitas posterior dari A pada kondisi B (*posterior probability).*

$P(B|A)$ = Probabilitas posterior dari B pada kondisi A (*likelihood).*

$P(A)$ = Probabilitas prior dari A (*class prior probability).*

$P(B)$ = Probabilitas prior dari B (*predictor prior probability).*

Proses untuk menghitung probabilitas kelas suatu data dimulai dengan menentukan *likelihood* berdasarkan dataset yang digunakan, menggunakan metode yang sesuai dengan bentuk dari data yang digunakan. *Likelihood* yang diperoleh akan dikalikan dengan probabilitas dari masing-masing kelas.

Hasil dari proses tersebut akan digunakan sebagai acuan untuk mengklasifikasi data baru. Pada praktiknya, seringkali P(B) dihiraukan, karena nilai P(B) selalu tetap.

2.2.3 *K-Nearest-Neighbor*

Algoritma *K-Nearest-Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Kedekatan didefinisikan dalam jarak metrik, seperti jarak *Euclidean*, jarak *Manhattan*, jarak *Minkowsky*, dan jarak *Chebychev*. Jarak tersebut dapat dicari dengan menggunakan persamaan berikut ini:

$$D_{euclidean} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \ldots(2)$$

$$D_{manhattan} = \sum_{i=1}^{n}|x_i - y_i| \quad \ldots(3)$$

$$D_{minkowsky} = \left(\sum_{i=1}^{n}|x_i - y_i{}^r|\right)^{1/r} \quad \ldots(4)$$

$$D_{chebychev} = \max_i |x_i - y_i| \quad \ldots(5)$$

2.2.4 *Logistic Regression*

Pada algoritma *Logistic Regression*, independen variabel didefinisikan sebagai persamaan regresi linear

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad \ldots(6)$$

Algoritma *Logistic Regression* memprediksi probabilitas keanggotaan variabel independen dalam suatu kelas menggunakan model fungsi logit transform atau invers logit dengan persamaan (Hadiwandra, 2019):

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \quad \ldots(7)$$

2.3 Pengukuran Kinerja Model Klasifikasi

Salah satu elemen yang dapat digunakan untuk mengukur kinerja model klasifikasi yaitu dengan pengukuran *predictive accuracy* dengan melihat nilai *accuracy* pada *Confusion Matrix*, kurva ROC dqan nilai AUC.

2.3.1 *Accuracy*

Pengukuran *Predictive accuracy* suatu model klasifikasi dilakukan berdasarkan pada perhitungan jumlah objek yang dapat di prediksi dengan benar dan jumlah objek yang tidak dapat diprediksi dengan benar. Perhitungan ini di tabulasikan kedalam tabel *Confusion Matrix*.

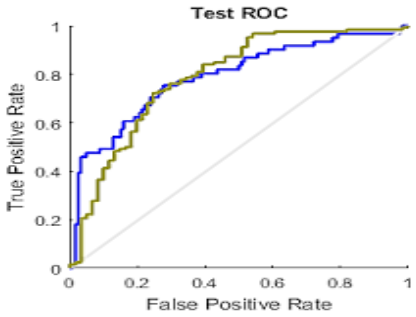
Tabel 2. Confusion Matrix

Prediksi	**Aktual**	
	Bukan *Phishing*	*Phishing*
Bukan *Phishing*	TP	FP
Phishing	FN	TN

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+FP+FN+TN} \quad \ldots(8)$$

Pada Tabel 2 merupakan bentuk *confusion matrix* hasil dari sebuah model yang dibangun, yaitu, *True positive* (TP) adalah nilai *class* Bukan *Phishing* yang diidentifikasikan benar, Kemudian, *true negative* (TN) adalah nilai *class Phishing* yang diidentifikasikan benar. *False positive* (FP) adalah nilai *class* Bukan *Phishing* yang diidentifikasikan salah, *False negative* (FN) adalah nilai *class Phishing* yang diidentifikasikan salah.

2.3.2 Kurva ROC



Gambar 1. Contoh kurva ROC

Perbandingan kinerja model klasifikasi juga dapat dilakukan secara visual menggunakan grafik kurva ROC (*Receiver Operating Characteristic*) (Gorunescu, 2011).

Pada kurva ROC terdapat garis diagonal yang dapat menunjukkan baik atau buruknya hasil klasifikasi suatu model. Titik yang berada diatas garis diagonal menunjukkan hasil yang baik dan titik yang berada dibawah menunjukkan hasil yang buruk. Suatu titik pada kurva ROC dikatakan lebih baik dari titik lain apabila mempunyai *True Positive rate* yang lebih tinggi dan *False Positive rate* yang lebih rendah.

2.3.3 Nilai AUC

Cara lain untuk membandingkan kinerja model klasifikasi adalah dengan menghitung luas area dibawah kuva ROC atau *Area Under The Curve* (AUC). Semakin besar nilai AUC semakin baik pula kinerja model klasifikasi. Tingkat akurasi dari suatu model klasifikasi dapat ditentukan berdasarkan kriteria nilai AUC sebagai berikut: 0.90 - 1.00 dikategorikan sempuna (*excellent classification*); 0.80 - 0.90 dikategorikan baik (*good classification*); 0.70 - 0.80 dikategorikan adil (*fair classification*); 0.60 - 0.70 = buruk (*poor classification*); dan 0.50 - 0.60 dikategorikan gagal (*failure*) (Gorunescu, 2011).

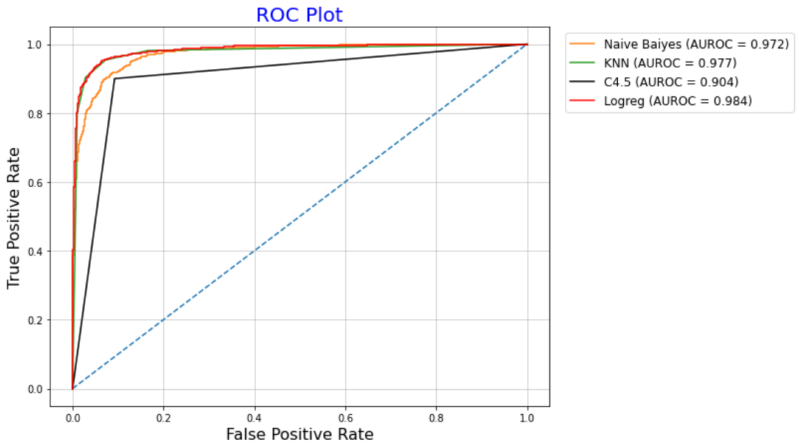
3. HASIL DAN PEMBAHASAN

Sebagaimana tujuan dari penelitian ini untuk mengetahui kelas model klasifikasi yang lebih baik dan relative optimal pada dataset phishing, maka dalam penelitian ini dilakukan pengujian dalam membandingkan kinerja algoritma dari beberapan kelas model yang berbeda dan menggunakan dataset phishing. Pengujian dilakukan menggunakan *python* versi 3.10.6 yang dijalankan diatas platform Ms. Windows 10.1 Enterprise 64bit dengan spesifikasi Intel® Core™ i3-1005G1 CPU @ 1.20GHz dan 4GB RAM. Pada penelitian ini digunakan empat algoritma yaitu algoritma *Decission Tree*, algoritma *Naïve Bayes*, algoritma *K-Nearest-Neighbor*, dan algoritma *Logistic Regression*. Evaluasi model dilakukan dengan melihat nilai *accuracy* dan nilai AUC yang dihasilkan.

Tabel 3. Hasil *Confusion Matrix*

Algoritma	TP	FP	FN	TN	ACC
Decission Tree	1553	158	172	1546	90,37%
Naive Bayes	1526	185	126	1592	90,93%
K-Nearest-Neighbor	1663	48	167	1551	93,72%
Logistic Regression	1625	86	112	1606	94,22%

Pada Tabel 2 terlihat hasil dari bentuk *confusion matrix* sebuah model yang dibangun untuk mengetahui kinerja berdasarkan nilai TP, FP, FN, daan TN. Nilai ACC pada Tabel 2 menunjukkan persentase akurasi dari empat model, terlihat pada tabel tersebut nilai ACC dari empat model di atas 90%, yang terkecil adalah *Decission Tree* dengan persentase akurasi sebesar 90,37% sedangkan *Logistic Regression* yang memiliki persentase 94,22% adalah yang terbesar dari keempat model tersebut.



Gambar 2. Kurva ROC

Pada Gambar 2 juga terlihat bagaimana kurva ROC dari *Logistic Regression* berada diatas kurva ROC algoritma lainnya. Jika dilihat dari nilai AUC nya, keempat algoritma termasuk model klasifikasi yang sempurna atau *excellent classification*. Dengan demikian untuk kasus ini algoritma *Logistic Regression* lebih optimal dibanding algoritma lainnya.

4. SIMPULAN DAN SARAN

Pengujian dilakukan dengan membagi dataset dengan ratio 70:30 menggunakan empat algoritma dari masing-masing metode klasifikasi, yaitu *Decission Tree, Naive Bayes, K-Nearest-Neighbor,* dan *Logistic Regression.* Hasil evaluasi menggunakan nilai *accuracy* dan nilai AUC menunjukkan bahwa model *Logistic Regression* merupakan model terbaik dengan nilai berturut-turut adalah 94,22% dan 99,4%. Penelitian ini masih terbatas pada perbandingan beberapa algoritma sederhana dari kelas model klasifikasi. Pengembangan selanjutnya dapat menngunakan algoritma lainnya yang beragam untuk mencari model klasifikasi yang terbaik.

DAFTAR PUSTAKA

Gorunescu, F. (2011). *Data Mining:Concepts,Models and Techniques* (Vol. 12). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642- 19721-5

Hadiwandra, T. Y. (2019). Perbandingan Kinerja Model Klasifikasi Decission Tree, Naive Bayes, K-Nearest-Neighbor, Logistic Regression, Rule Base pada 4 Dataset Berbeda. *Sains dan Teknologi Informasi*, *5*(1), 70-78

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining:Concepts and Techniques* (3rd ed.). Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset. Diakses tanggal : 27 September 2022. Pukul 19.15.