“BIMASAKTI”
Jurnal Riset Mahasiswa Bidang Teknologi Informasi

Volume 6 Nomor 2
e-issn : 2355-4401
Available online at https://ejournal.unikama.ac.id/index.php/JFTI

INVESTIGASI PERBANDINGAN *COSINE SIMILARITY* DAN *EUCLIDEAN DISTANCE* DALAM DETEKSI *PHISHING ATTACK* MENGGUNAKANAN METODE K-NEAREST NEIGHBOR

Bibiana Da Frosa[1)], **Akhmad Zaini**[2)], **Muhammad Priyono Tri**[3)]

Universitas PGRI Kanjuruhan Malang[1),2),3)]
bibianadafrosa16@gmail.com

Abstrak

Perkembangan teknologi informasi mempengaruhi berbagai aspek kehidupan. Namun, dampak positif ini juga membuka celah bagi kejahatan dunia maya yang semakin berkembang, dikenal sebagai cybercrime (Iman et al., 2020). Kejahatan ini, seperti carding, hacking, dan phishing, mengancam keamanan di ranah digital (Gulo et al., 2021). Phishing, sebagai salah satu bentuk cybercrime, melibatkan pengiriman link palsu untuk mencuri informasi korban (Wibowo & Fatimah, 2017). Di tengah perkembangan teknologi sistem informasi, data mining muncul sebagai solusi, memungkinkan penggalian informasi berharga dari data besar. K-Nearest Neighbors (KNN) adalah salah satu algoritma machine learning yang digunakan untuk klasifikasi dan regresi (Dewi Obert & Gusmana, 2018). Dalam K-Nearest Neighbor, metode jarak seperti *euclidean distance, Manhattan distance, cosine similarity, dan jaccard similarity* umum digunakan. Fokus penelitian ini adalah pada *euclidean distance* dan *cosine similarity* yang dianggap efisien dan umum digunakan. Hasil evaluasi menunjukkan bahwa kedua metode, *cosine similarity* dan *Euclidean distance*, memiliki tingkat akurasi dan kecepatan yang serupa dalam deteksi serangan *phishing*. Namun, *Euclidean distance* menonjol dalam deteksi *phishing* dengan tingkat akurasi sebesar 87.70% dan kecepatan 0.0172. Sementara itu, *cosine similarity* mencapai tingkat akurasi 87.57% dengan kecepatan 0.0360. Analisis looping menegaskan keunggulan kecepatan *Euclidean distance* secara konsisten. Dalam deteksi *phishing attack*, *Euclidean distance* terbukti lebih efektif dalam akurasi dan kecepatan.

Kata Kunci : *K-Nearest Neighbor algorithm*; *Cosine Similarity*; *Euclidean Distance*

Abstract

The development of information technology affects various aspects of life. However, this positive impact also opens up opportunities for growing cybercrime, known as cybercrime (Iman et al., 2020). These crimes, such as carding, hacking, and phishing, threaten security in the digital realm (Gulo et al., 2021). Phishing, as a form of cybercrime, involves sending fake links to steal victim information (Wibowo & Fatimah, 2017). In the midst of the development of information technology systems, data mining has emerged as a solution, enabling all valuable information from big data. K-Nearest Neighbors (KNN) is a machine learning algorithm used for classification and regression (Dewi Obert & Gusmana, 2018). In K-Nearest Neighbor, distance methods such as euclidean distance, Manhattan distance, cosine similarity, and jaccard similarity are commonly used. The focus of this research is on euclidean distance and cosine similarity which are considered efficient and commonly used. The evaluation results

show that the second method, cosine similarity and Euclidean distance, has a similar level of accuracy and speed in detecting phishing attacks. However, Euclidean distance stands out in phishing detection with an accuracy rate of 87.70% and a speed of 0.0172. Meanwhile, cosine similarity reaches an accuracy rate of 87.57% with a speed of 0.0360. Looping analysis consistently confirms the Euclidean distance speed advantage. In phishing attack detection, Euclidean distance is proven to be more effective in accuracy and speed.

Keywords **:** *K-Nearest Neighbor algorithm*; *Cosine Similarity*; *Euclidean Distance*

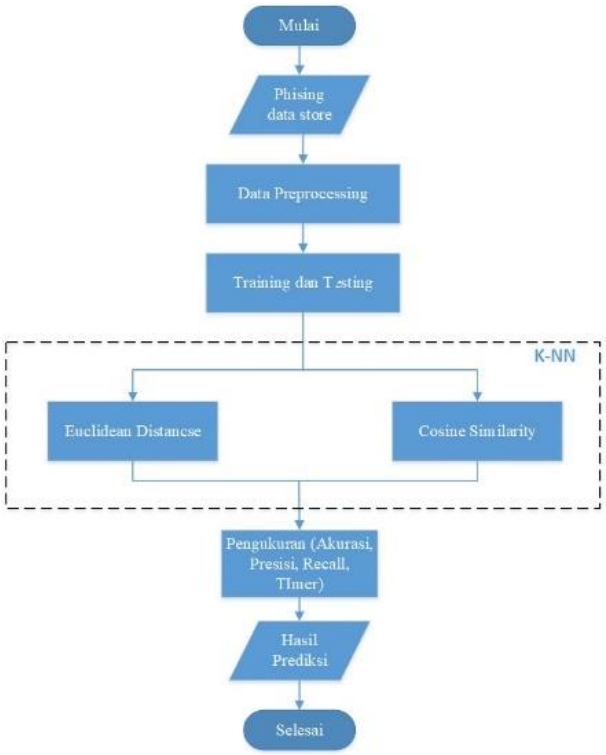
1. PENDAHULUAN

Teknologi informasi mampu mengubah realitas ekonomi, budaya, politik, dan hukum. Seiring berkembangnya teknologi informasi mampu memberikan dampak positif bagi banyak orang namun hal ini juga menyebabkan munculnya kejahatan-kejahatan baru yang disebut dengan kejahatan dunia maya baru melalui jaringan internet. Dimana terdapat beberapa orang yang memanfaatkan celah keamanan pada teknologi informasi pada jaringan internet sebagai sarana untuk melakukan kejahatan yang selanjutnya dikenal dengan cybercrime (Iman et al., 2020).
Cybercrime merupakan fenomena yang sangat mengkhawatirkan, mengingat tindakan carding, hacking, penipuan, terorisme, dan penyebaran informasi yang mengganggu menjadi bagian dari aktivitas pelaku *cybercrime* (Gulo et al., 2021). Banyak jenis dan ragam *cybercrime* salah satunya phishing atau dikenal sebagai carding. *Phishing* merupahkan penipuan atau serangan di mana penyerang mengirimkan *link* palsu dalam upaya untuk mencuri informasi rahasia korban (Wibowo & Fatimah, 2017).
Perkembangan teknologi system informasi telah memecahkan banyak masalah di berbagai bidang, salah satunya adalah penerapan data mining. Data mining adalah proses pengalian informasi atau pola yang berguna dan bermakna dari kumpulan data yang besar dan kompleks. K-Nearest Neighbors (KNN) adalah salah satu metode dalam pembelajaran mesin yang digunakan untuk klasifikasi dan regresi (Dewi Obert, & Gusmana, 2018). Ketika menggunakan algoritma K-Nearest Neighbors (KNN) untuk klasifikasi data, pilihan jenis jarak yang digunakan dapat mempengaruhi performa dan hasil dari model yang dibangun. Beberapa jenis jarak yang umum digunakan dalam K-Nearest Neighbor antara lain: *euclidean distance, Manhattan distance, cosine similarity, dan jaccard similarity*. Beberapa jenis jarak yang biasa dipakai untuk metode K-Nearest Neighbor ini adalah *euclidean distance* dan *cosine similarity*. Dari semua pernyataan tersebut alasan penelitian ini lebih memilih *euclidean distance* dan *cosine similarity* merupahkan dua algoritma yang umum yang digunakan dan memiliki performa yang efisien.
Penelitian yang dilakukan oleh (Bagus Trianto et al., 2022). Tentang Kombinasi K-Nearest Neighbor dan cosine similarity untuk perbandingan prediksi serangan firewall pada jaringan komputer. Dari penelitian tersebut disajikan hasil analisis bahwa algoritma perhitungan cosine similarity dengan model K-Nearest Neighbor dapat memberikan performa untuk prediksi serangan firewall dibandingkan dengan perhitungan jarak lainnya. Hal ini dapat dibuktikan pada tingkat akurasi dan hasil yang sangat baik, dengan adanya pembuktian tersebut maka kesimpulannya bahwa perhitungan cosine similarity ini menjadi acuan terbaik.
Terdapat juga penelitian lainya yang membandingkan perhitungan Euclidean distance dengan model K-Nearest Neighbor seperti penelitian yang dilakukan oleh (Wahyono et al., 2020). Pada penelitian tersebut membandingkan tingkat akurasi dari perhitungan jarak pada metode K-Nearest Neighbor dalam kasus klasifikasi data tekstual. Hasilnya tingkat akurasi pada perhitungan Euclidean distance lebih tinggi dibandingkan dengan perhitungan jarak lainnya hal

ini membuktikan bahwa *Euclidean distance* juga menjadi acuan terbaik dibandingkan yang lainnya. Dengan memperhatikan hal sudah dijabarkan pada latar belakang diatas, penelitian ini bertujuan untuk membandingkan *cosine Similarity* dan *Euclidean distance* dengan metode K-Nearest Neighbor, untuk menganalisis kelayakan tingkat akurasi dan proses kecepatan, perhitungan *Cosine similarity* lebih baik dari pada *Euclidean distance* atau sebaliknya.

2. METODE / ALGORITMA

Pada penelitian ini menggunakan metode penelitian kuantitatif, yaitu suatu metode penelitian yang menggunakan data berupa angka untuk kemudian diolah dan dianalisis untuk mendapatkan suatu informasi ilmiah dibalik angka tersebut. Data yang digunakan adalah data *Phishing URL Detection*. Dalam penelitian ini terdapat beberapa tahap dalam pengerjaan untuk digunakan sebagai awal dalam menyelesaikan penelitian. Tahapan yang akan dikerjakan meliputi pengumpulan dataset setelahnya dilanjutkan preprocessing. Pada tahap *prepocessing* ini dilakukan proses pembobotan kata dalam setiap dokumen melalui tahap Tf-IDF, setalah proses TF-IDF akan menghasilkan data intensi baru dimana data ini akan langsung dikelola oleh cosine similarity dan Euclidean distance dengan menggunkana metode *K-Nearest Negihbor*, setalah mendapat tingkat nilai akurasi dan kecepatan dalam perhitungannya maka dilanjutkan proses prediksi dalam prediksi ini terdapat dua label yaitu *bad* dan *good* dimana *bad* ini merupahkan data *phishing* dan *good* merupahkan data normal.



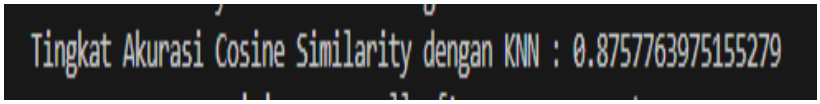
Gambar .1 Metode penelitian

3. HASIL DAN PEMBAHASAN

A. Mengihtung akurasi dan kecepatan *Cosine Similarity* menggunakan metode K-Nearest Neighbor.

Pada penelitian ini algoritma K-Nearest Neighbor dilibatkan dengan metiks *cosine similarity* untuk melakukan prediksi pada data testing serta evaluasi performa model. Evaluasi ini penting

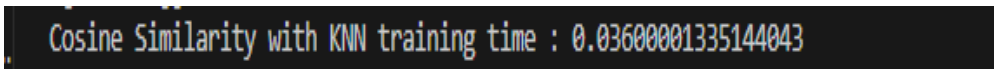
untuk memahami beberapa baik model dapat melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya. Hasil dari tingkat akurasi K-Nearest Neighbor dengan *cosine similarity*, dengan adanya perhitungan melalui evaluasi model tersebut akhirnya mendapatkan tingkat akurasi berupa 87.57%.



Gambar. 2 Hasil akurasi consine similarity dengan KNN

Meenghitung kecepatan dalam sekali

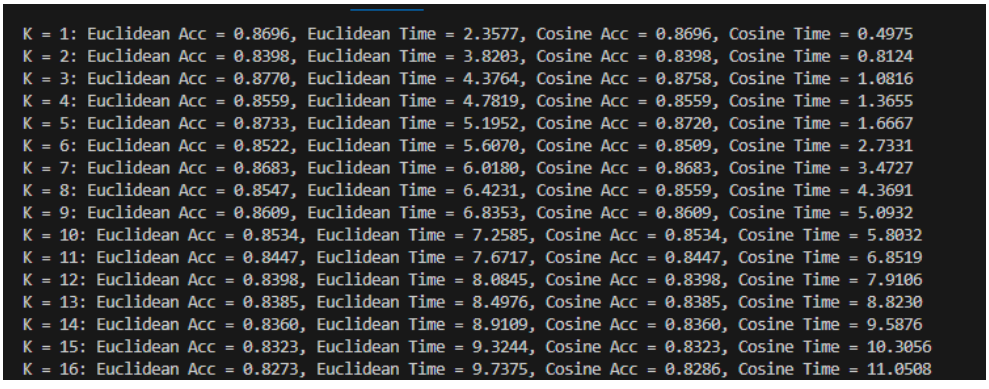
Perhitungan kecepatan (speedup) dalam konteks penghitungan waktu pada proses K-Nearest Neighbor dengan *cosine similarity* bisa merujuk pada perbandingan waktu yang dibutuhkan untuk menjalankan proses tersebut dengan menggunakan berbagai teknik atau pendekatan yang berbeda. Hasil dari perhitungan kecepatan *cosine similarity* dalam sekali yang menunjukan hasil dengan 0.0360 ini membiktikan bahwa perhitungan kecepatan cosine similarity masih lamban.



Gambar .3 Perhitungan kecepatan cosine similarity dengan KNN

Menghitung akurasi kecepatan perlooping (1-100)

Perhitungan looping dalam menghitung timer pada K-Nearest Neighbor dengan cosine similarity bisa merujuk pada penggunaan konstruksi loop (perulangan) untuk melakukan beberapa iterasi dari proses yang sama, seperti pelatihan model atau prediksi, dan kemudian mengukur waktu yang diperlukan untuk setiap iterasi tersebut. Hasil dari perhitungan kecepatan *cosine similarity* menggunakan looping dari perhitungan yang kedua ini juga menghsilkan bahwa *cosine similarity* ini masih lamban dalam hal perhitungan mencari nilai kecepatan bisa pada gambar dibawah ini.



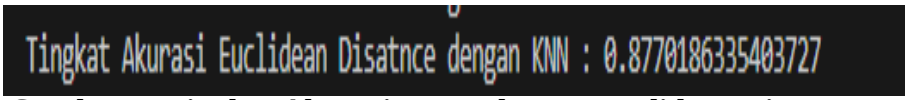
Gambar .4 Perhitungan kecepatan cosine similarity menggunakan looping

B. Mengihtung akurasi dan kecepatan *Euclidean Distance* menggunakan metode K-Nearest Neighbor.

Menghitung tingkat akurasi akurasi

Pada penelitian ini perhitungan nilai akurasi dalam *Euclidean Distance* merujuk pada evaluasi performa model *machine learning*, seperti K-Nearest Neighbors (KNN), yang menggunakan metrik jarak *Euclidean* untuk mengukur jarak antara titik data dalam ruang fitur. Dari

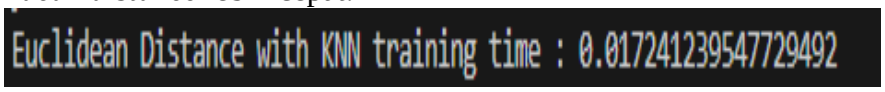
perhitungan melalui model tersebut dapat diketahui hasil dari tingkat akurasi perhitungan k-Nearest Neighbor dengan *Euclidean distance* mendaptakan hasil yaitu 87.70%.



Gambar .5 Tingkat Akurasi K-NN dengan Euclidean Distance

Meenghitung kecepatan dalam sekali

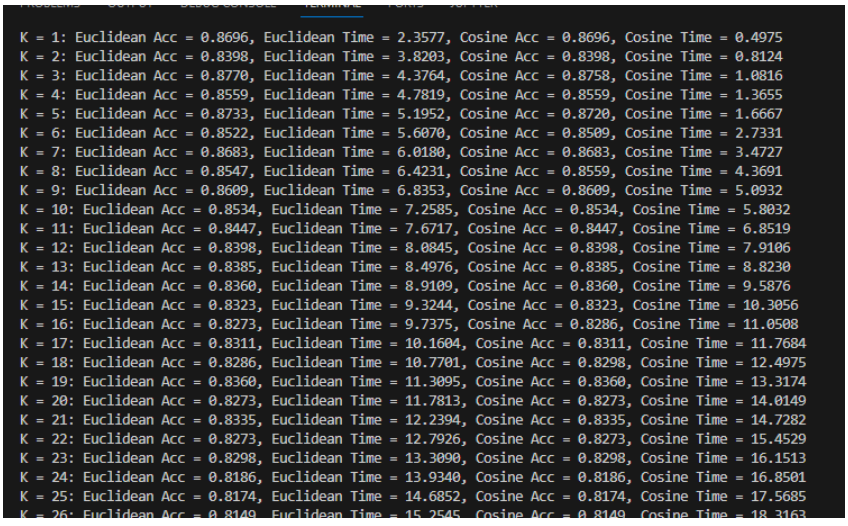
Perhitungan kecepatan (*speedup*) dalam konteks penghitungan waktu pada proses K-Nearest Neighbor dengan *Euclidean distance* bisa merujuk pada perbandingan waktu yang dibutuhkan untuk menjalankan proses tersebut dengan menggunakan berbagai teknik atau pendekatan yang berbeda. Hasil dari perhitungan kecepatan *Euclidean distance* yang menunjukan hasil yaitu 0.0172 dengan melihat hasil tersebut, dengan ini membuktikan bahwa perhitungan dalam kecepatan *Euclidean distance* lebih cepat.



Gambar .6 Perhitungan kecepatan Euclidean distance

Menghitung akurasi kecepatan perlooping (1-100)

Pada perhitungan kecepatan (speedup) menggunakan perulangan (looping) dalam *Euclidean Distance* pada K-Nearest Neighbors (KNN), perhitungan kecepatan sering kali melibatkan perbandingan waktu yang dibutuhkan untuk proses yang sama dengan variasi jumlah tetangga terdekat (K) atau variasi metrik jarak. Hasil dari perhitungan ke dua yang berupa looping mendapatkan hasil bahwa perhitungan kecepatan dalam kasus *phishing attack* ini membuktikan kecepatan dari *Euclidian distance* lebih cepat bisa dilihat pada gambar dibawah ini.



Gambar .7 Perhitungan kecepatan *Euclidean distance* berupa looping

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan mengenai Investigasi perbandingan *Cosine Similarity* Dan *Euclidean Distance* Dalam Deteksi *Phishing Attack* menggunakan metode K-Nearest Neighbor maka dapat di ambil kesimpulan yaitu Pengujian yang dilakukan

menghasilkan tingkat akurasi rata-rata sebesar 87.70% dengan kecepatannya yaitu 0.0172. Berdasarkan analisis jumlah data, algoritma *Euclidean distance* masih menunjukkan keunggulan dalam mendeteksi *phishing attack* pada *URL* menggunakan alat *visual Studyo Code* dengan bahasa pemograman *Python*, meskipun perbedaannya hanya sedikit.
Saran untuk penelitian di masa yang akan datang Investigasi perbandingan *Cosine Similarity* Dan *Euclidean Distance* Dalam Deteksi *Phishing Attack* menggunakan metode K-Nearest Neighbor ini yaitu diperlukan pengujian lebih lanjut dengan menggunakan metode lain seperti Naive Bayes guna memperoleh pemahaman mendalam mengenai metode perhitungan yang paling efektif untuk diterapkan dalam deteksi phising *attack*.

5. REFERENSI

[1] Bagus Trianto, R., Triyono, A., Malita Puspita Arum, D., Komputer, I., Sains dan Kesehatan, F., An Nuur, U., Gajah Mada No, J., Purwodadi, K., & Grobogan, K. (2022). Kombinasi Metode K-Nearest Neighbor dengan Cosine Similarity untuk Prediksi Serangan Firewall pada Jaringan Komputer. *Jurnal Informatika UniversitasPamulang*, *6*(4), 672–679. http://openjournal.unpam.ac.id/index.php/informatika/article/view/12680

[2] Gulo, A. S., Lasmadi, S., & Nawawi, K. (2021). Cyber Crime dalam Bentuk Phising Berdasarkan Undang-Undang Informasi dan Transaksi Elektronik. *PAMPAS: Journal of Criminal Law*, *1*(2), 68–81. https://doi.org/10.22437/pampas.v1i2.9574

[3] Iman, N., Susanto, A., & Inggi, R. (2020). Analisa Perkembangan Digital Forensik dalam Penyelidikan Cybercrime di Indonesia (Systematic Review). *Jurnal Telekomunikasi Dan Komputer*, *9*(3), 186. https://doi.org/10.22441/incomtech.v9i3.7210

[4] Wahyono, W., Trisna, I. N. P., Sariwening, S. L., Fajar, M., & Wijayanto, D. (2020). Comparison of distance measurement on k-nearest neighbour in textual data classification. *Jurnal Teknologi Dan Sistem Komputer*, *8*(1), 54–58. https://doi.org/10.14710/jtsiskom.8.1.2020.54-58

[5] Wibowo, M. H., & Fatimah, N. (2017). Ancaman Phishing Terhadap Pengguna Sosial Media Dalam Dunia Cyber Crime. *JOEICT(Jurnal of Education and Information Communication Technology)*, *1*(1), 1–5. https://www.jurnal.stkippgritulungagung.ac.id/index.php/joeict/article/view/69

[6] Dewi, R. F. K., Obert, & Gusmana, R. (2018). Implementasi Metode K-Nearest Neighbor (KNN) dalam Pengelompokan Status Ekonomi Warga. Journal of Big Data Analytic and Artificial Intelligence, 4(1), 15–22.