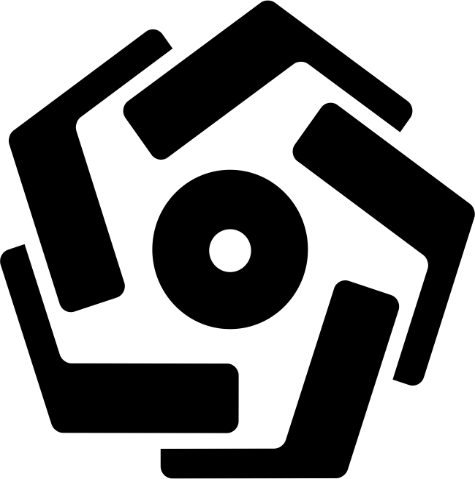
**SELEKSI FITUR BERBASIS INFORMATION GAIN PADA KINERJA ALGORITMA KLASIFIKASI PADA DETEKSI WEB PHISHING**

**SKRIPSI**



disusun oleh

Anggun Wahyu Andriyanto

17.11.1732

**PROGRAM SARJANA**

**PROGRAM STUDI INFORMATIKA**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**YOGYAKARTA**

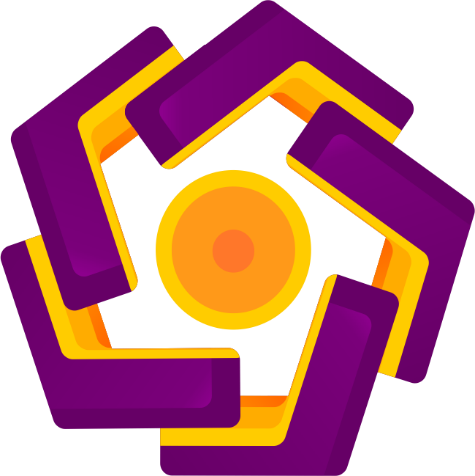
**2020**

**SELEKSI FITUR BERBASIS INFORMATION GAIN PADA KINERJA ALGORITMA KLASIFIKASI PADA DETEKSI WEB PHISHING**

**SKRIPSI**

untuk memenuhi sebagian persyaratan   
mencapai gelar Sarjana

pada Program Studi Sistem Informasi



disusun oleh

Anggun Wahyu Andriyanto

17.11.1732

**PROGRAM SARJANA**

**PROGRAM STUDI SISTEM INFORMASI**

**SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER**

**AMIKOM YOGYAKARTA**

**YOGYAKARTA**

**2020**

**BAB I**  
Pendahuluan

1. 1. **Latar Belakang**

Perkembangan internet dari tahun ke tahun membuat fasilitas yang ditawarkannya menjadi lebih banyak. Dari segi kemudahan akses hingga banyaknya fitur yang ditawarkan pastinya akan mudah membuat orang-orang tertarik untuk menggunakannya, sehingga tidak dipungkiri jika penggunanya tidaklah sedikit. Sebagai contohnya terdapat media sosial, internet atau mobile banking. Dua hal tersebut telah memberikan kemudahan bagi orang-orang untuk saling berhubungan meski dipisahkan oleh jarak dan membantu proses transaksi, yang mana bisa diakses dan dilakukan dari mana saja secara onlinedan *realtime*. Namun dibalik kemudahan tersebut, bisa menjadi celah keamanan bagi pengguna internet dalam bertransaksi secara online maupun mengakses media sosial seperti data pribadi, kata sandi, e-mail, informasi sensitif internet banking atau mobile banking. Hal tersebut biasanya dilakukan oleh pelaku dengan bantuan menggunakan situs phising. Situs Phising merupakan suatu metode untuk melakukan penipuan dengan mengelabui target dengan maksud untuk mencuri akun target. Istilah ini berasal dari kata “fishing” = “memancing” korban untuk terperangkap dijebakannya. Sedangkan situs phising adalah situs yang dibuat semirip mungkin dengan situs aslinya dan digunakan untuk mengelabui lalu menjebak korban yang mengaksesnya [1]. Phising juga terjadi pada media sosial seperti Facebook, Gmail, dan Twitter. Dimana situs phising yang dibuat akan memberikan informasi layaknya situs aslinya kemudian ketika informasi pengguna dimasukkan pada kolom *login,* situs tersebut akan mengarah ke situs phising daripada ke situs aslinya [2]. Dari laporan APWG (*Anti-Phishing Working Group*) [3], tahun 2019 menjadi naik turunnya perbuatan phishing, meskipun begitu tetap saja terjadi peningkatan pesat pada bulan-bulan tertentu. Dari laporan APWG kuartal keempat 2019, pada bulan Oktober – Desember ada 162.555 situs web yang terdeteksi oleh APWG. Hasil tersebut turun dari 266.387 pada bulan Juli – September.

Hal tersebut bisa menimbulkan ketidakpercayaan pengguna internet pada situs media sosial maupun situs transaksi. Oleh sebab itu dibutuhan sistem yang mampu mendeteksi situs phishing untuk mencegah kriminalitas phishing sehingga mengurangi dan bisa menghindari kerugian situs phishing terhadap pengguna internet. Dari beberapa penelitian sebelumnya digunakan metode klasifikasi untuk deteksi situs phishing dan non-phishing sehingga dapat dilakukan pencegahan terkena serangan phishing. Metode klasifikasi menggunakan semua fitur yang terdapat pada sebuah data untuk membangun model, namun menurut [4] tidak semua fitur tersebut relevan terhadap hasil klasifikasi. Apabila hal tersebut pada data yang memiliki ukuran besar, maka kinerja klasifikasi akan menjadi tidak efisien dan efektif, misalkan beban yang dilakukan oleh komputer untuk melakukan komputasi menjadi lebih berat, waktu pemrosesan menjadi lebih lama.

Kemudian diterapkanlah seleksi fitur berbasis Information Gain ----, yang berguna untuk menyeleksi fitur-fitur yang relevan pada dataset. Fitur yang memiliki relevansi rendah akan dibuang sehingga mengurangi beban pemrosesan pada kinerja metode klasifikasi.

Oleh karena itu pada penelitian ini akan diterapkan seleksi atribut yang berbasis Information Gain terhadap kinerja algoritma klasifikasi yang popular yaitu Decision Tree, Naïve Bayes dan K-NN. Dengan demikian kinerja klasifikasi yang dihasilkan dapat terukur secara sistematis setelah melalui seleksi atribut atau seleksi fitur, sehingga hasil penelitian ini dapat digunakan untuk menunjang penelitian lain yang sejenis dikemudian hari. Tentunya apabila model klasifikasi dari penelitian ini diimplementasikan, maka dapat mengurangi resiko pengguna internet terkena serangan phishing.

* 1. **Rumusan Masalah**

Berdasarkan latar belakang yang telah disampaikan, di bawah ini adalah rumusan masalah atau sesuatu yang menjadi pertanyaan-pertanyaan yang harus diselesaikan oleh peniliti pada penelitian yang akan dibuat :

1. Apakah seleksi fitur mempengaruhi kinerja algoritma klasifikasi?
2. Fitur-fitur apa saja yang relevan terhadap hasil klasifikasi?
3. Adakah peningkatan kinerja klasifikasi setelah dilakukan seleksi atribut?
4. Adakah penurunan kinerja klasifikasi setelah dilakukan seleksi atribut?
   1. **Batasan Masalah**

Untuk membatasi permasalahan yang lebih luas pada konteks objek, maka ruang lingkup dalam penelitian ini akan dibatasi sesuai kebutuhan dan kemampuan. Batasan masalah dalam penelitian ini antara lain:

1. Data situs phishing didapatkan dari repository UCI Machine Learning : Phishing Websites Data Set dan PhishThank.
2. Dalam penelitian ini akan digunakan Information Gain untuk seleksi atribut atau fitur pada dataset kemudian dilakukan klasifikasi menggunakan algoritma klasifikasi. Dengan batas hasil akurasi setelah seleksi atribut.
3. Metode klasifikasi yang digunakan dalam penelitian ini adalah Decision Tree, Naïve Bayes dan K-NN. Dengan batas hasil akurasi dan hasil penerapan
   1. **Maksud dan Tujuan Penelitian**

Di bawah ini adalah tujuan dari penelitian ini, yaitu:

1. Tujuan dari penelitian ini adalah menerapkan seleksi fitur berbasis Information Gain pada dataset latih, kemudian dilakukan klasifikasi menggunakan algoritma klasifikasi.
2. Tujuan dari penelitian ini adalah untuk mengetahui kinerja klasifikasi setelah diterapkan seleksi atribut yang relevan menggunakan Information Gain, apakah meningkat atau menurun.
3. Pemilihan algoritma Decision Tre, Naïve bayes, K-NN dilakukan untuk pembanding manakah yang hasilnya paling akurat.
   1. **Manfaat Penelitian**

Adapun manfaat yang diharapkan dari penelitan ini adalah:

1. Pemrosesan klasifikasi pada deteksi web phishing menjadi lebih efektif dan efisien setelah dilakukan seleksi atribut.
2. Mengetahui atribut atau fitur apa saja yang relevan pada data set.
3. Mengetahui pengaruh penerapan seleksi fitur pada kinerja klasifikasi.
   1. **Metode Penelitian**

Dalam penelitian ini, peneliti menggunakan data yang berasal dari repository UCI Machine Learning bagian Phishing Websites Data Set. Data tersebut yang nantinya akan digunakan untuk menguji ketiga metode klasifikasi yang sudah dilakukan seleksi atribut. Sehingga hasil uji tersebut bisa dibandingkan mana yang memiliki hasil akurasi tertinggi dan juga berapa peningkatan yang terjadi pada ketiga klasifikasi tersebut.

1. **Sistematika Penulisan**

Pada bagian ini dituliskan urut-urutan dan sistematika penulisan yang dilakukan. Berikan ringkasan mengenai isi masing-masing bab.

* **BAB I PENDAHULUAN**

Pada bab ini dijelaskan mengenai latar belakang, rumusan masalah, batasan masalah, maksud dan tujuan penelitian, manfaat penelitian, metode penelitan, dan sistematika penulisan dalam penelitan.

* **BAB II LANDASAN TEORI**

Pada bab ini akan dijelaskan mengenai landasan teori-teori dan kajian pustaka dari berbagai penelitian yang memiliki keterkaitan dengan penelitian ini. Hal tersebut berguna untuk memperkuat dasar, analisa, penulisan dan alasan dilakukannya penelitian ini. Sumber dari landasar teori ini juga berasal dari buku, jurnal, blog, maupun laporan yang berasal dari internet.

* **BAB III METODOLOGI PENELITIAN**

Pada bab ini akan dijelaskan mengenai langkah-langkah penelitian beserta metode yang digunakan.

* **BAB IV HASIL DAN PEMBAHASAN**

Pada bab ini akan dilakukan perancangan sistem serta pembahasannya. Kemudian di bab ini juga hasil penelitian akan didapatkan dan dilakukan pembahasannya.

* **BAB V KESIMPULAN**

Pada bab ini berisi kesimpulan dari penelitian ini dan juga saran bagi penelitian mendatang yang berasal dari kekurangan dari penelitian ini.

Daftar pustaka

1. Y. Ping, G. Yuxiang, Z. Futai, Y. Yao, W. Wei, Z. Ting. “Web Phishing Detection Using a Deep Learning Framework”, Wireless Communication and Mobile Computing, Vol. 2018, Article ID 4678746, 9 pages, 2018. https://doi.org/10.1155/2018/4678746
2. G. Varshney, M. Misra, P. K. Atrey, “A Survey and Classification of Web Phishing Detection Schemes”, Wiley Online Library., Vol. 9, No. 18, hal. 6266-6284, Okt 2016.
3. APWG, “Phishing Activity Trends Report, 4th Quarter 2019.” 2020.
4. H. N. Firqiani, A. Kustiyo, E. P. Giri “Seleksi Fitur Menggunakan Fast Correlation Based Filter Pada Algoritma Voting Feature Interval 5”, Institut Pertanian Bogor, 2008.
5. N. Abdelhamid, A. Ayesh, F. Thabtah “Phishing Detection Based Associative Classification Data Mining.”, ELSEVIER, Vol. 41, Issue 13, hal 5498-5959, Okt 2014.