BAB I
PENDAHULUAN

1.1 Latar Belakang

Dunia pada saat ini tengah mengalami perkembangan pada bidang teknologi dan komunikasi secara massif, terlebih pada masa pandemi saat ini yang mengharuskan kita semua belajar bahkan bekerja secara daring. Semua teknologi yang menggunakan internetlah yang berkembang paling pesat yang semakin disempurnakan dengan kehadiran pandemi. Internet sudah seperti menjadi nafas baru kehidupan ditengah kehadiran pandemi. Perkembangan teknologi pasti diikuti dengan peningkatan penggunaan internet. Di Indonesia saja, selama pandemi Covid-19, penggunaan internet naik hingga 40%. Peningkatan inilah yang memicu banyaknya kejahatan kejahatan di dunia internet. Salah satu nya adalah penipuan. Salah satu teknik penipuan yang ada di internet adalah phising. Phising merupakan suatu metode untuk melakukan penipuan dengan mengelabui target dengan maksud untuk mencuri akun target [7]. Pada umumnya penjahat internet akan mencuri informasi berupa username, password dan data identitas sensitif lainnya, dari akun media sosial maupun akun nomor kartu kredit. Untuk melakukan hal ini, penjahat internet biasanya menggunakan web phising sebagai alat bantunya.

Web phising yaitu situs web yang dirancang oleh penjahat internet semirip mungkin dengan situs aslinya dari segi tampilan, konten, URL domain atau lainnya untuk mengelabui korbannya (pengguna internet) dengan membuat korban seolah-olah sedang mengakses halaman situs dari sumber yang sah [9]. Selain itu, ada juga situs phising yang didesain untuk memberikan informasi yang menyesatkan. Jika korban berhasil dikelabuhi dan memasukkan informasi yang sensitif yang diminta,

maka penjahat internet dengan mudah dapat menggunakan informasi tersebut pada situs yang sah untuk melakukan aktivitas yang tidak diinginkan dan tentunya akan menimbulkan kerugian bagi para korbannya baik data pribadi maupun kerugian finansial. Dari laporan APWG (*Anti-Phishing Working Group*), mulai pertengahan Maret, penjahat dunia maya diluncurkan berbagai phishing bertema COVID-19 dan serangan malware terhadap pekerja, layanan kesehatan fasilitas, dan yang baru saja menganggur. Jumlah situs phising yang terdeteki pada kuarta pertama tahun 2020 adalah 165.772, naik dari 162.155 pada kuartal keempat tahun 2019 [4]. Setelah dua kali lipat pada tahun 2020, jumlah phising menurun selama kuartal pertama tahun 2021. Namun, Januari 2021 adalah rekor tertinggi APWG, dengan 245.771 serangan yang belum pernah terjadi sebelumnya dalam satu bulan. Sektor lembaga keuangan, webmail, dan media sosial adalah yang paling sering menjadi korban phising di kuartal ini [3].

Untuk mengatasi maraknya phising yang terjadi di dunia maya, maka diperlukan sistem untuk mendeteksi situs phising agar bisa menghindari dan mengurangi kerugian pengguna internet. Dari beberapa penelitian sebelumnya, untuk deteksi web phising dan non-phising digunakan seleksi fitur dan metode klasifikai. Metode yang digunakan diantaranya yaitu random forest dan logistic regression. Pada penelitian [8] dilakukan deteksi web phising menggunakan metode Binary logistic regression dan menggunakan seleksi atribut yaitu correlation-based feature selection (CFS) dengan hasil akurasi yang tinggi. Pada penelitian [1] dilakukan perbandingan algoritma klasifikasi untuk menggidentifikasi web phising salah satunya yaitu Random Forest. Pada penelitian ini, nilai akurasi random forest masih dibawah algoritma Multilayer Perceptron.

Sedangkan pada pnelitian [18] nilai akurasi random forest dengan menggunakan seleksi fitur Spearman dan MICe TICe lebih tinggi daripada Logistic Regression.

Berdasarkan uraian diatas, peneliti memutuskan menggunakan random forest dan logistic regression untuk deteksi web phising meggunakan seleksi fitur berbasis korelasi atau correlation-based feature selection (CFS). Metode ini digunakan peneliti karena random forest dapat meningkatkan akurasi dan dapat menangani input variabel yang besar, menyeimbangkan error dalam unbalanced dataset [13]. Dan untuk logistic regression cocok digunakan untuk memprediksi keanggotaan variabel independen (prediktor) dalam dua group [30]. Sedangkan untuk CFS digunakan untuk menghapus atribut yang tidak relavan dan redundan. Dengan demikian peneliti memutuskan memilih judul **"Perbandingan Logistic Regression dan Random Forest menggunakan Correlation-Based Feature Selection Untuk Deteksi Website Phising".**

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka permasalahan yang dapat dirumuskan adalah :

1. Bagaimana hasil perbandingan antara algoritma Logistic Regression dan Random Forest?
2. Apakah penerapan seleks fitur mempengaruhi hasil kinerja algoritma Logistic Regression dan Random Forest?
3. Fitur apa saja yang saling berkorelasi dan relevan terhadap hasil klasifikasi?
4. Adakah peningkatan atau penurunan kinerja Logistic Regression dan Random Forest setelah dilakukan seleksi fitur?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini sebagai berikut:

1. Dalam penelitian ini akan membandingkan 2 algoritma yaitu logistic regression dan random forest.
2. Dataset yang digunakan diambil dari UCI *Machine Learning Repository*.
3. Dalam penelitian ini menggunakan Correlation-Based untuk seleksi fitur pada dataset.

1.4 Maksud dan Tujuan Penelitian

Maksud dan tujuan dalam pembutan penelitian ini, yaitu :

1. Membandingkan 2 algoritma untuk mengetahui manakah algoritma yang lebih akurat untuk deteksi web phising.
2. Mengukur kinerja algoritma logistic regression dan random forest setelah diterapkan seleksi fitur Correlation-Based.

1.5 Metode Penelitian

1.5.1 Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan pada penelitian ini, yaitu:

1. Data *training* dan *test* pada penelitian ini diperoleh dari UCI *Machine Learning Repository*.
2. Data situs phishing didapatkan dari PhishTank.
3. Data situs non-phishing (situs otentik) didapatkan dari Alexa dan Moz Top 500.

1.5.2 Tahapan Penelitian

Tahapan penelitian dimulai dengan melakukan studi literatur dari penelitian terkait. Kemudian peneliti mengambil dataset dari UCI *Machine Learning Repository*. Selanjutnya seleksi fitur diterapkan pada dataset tersebut untuk menyeleksi fitur yang sangat berkorelasi dengan (prediktif) kelas, namun tidak berkorelasi (tidak prediktif) satu sama lain. Jika fitur memiliki relevansi rendah atau bahkan tidak memiliki relevansi maka tidak akan dipakai. Hasil dari seleksi tersebut nantinya akan digunakan untuk menguji metode klasifikasi random forest dan logistic regression. Kemudian membandingkan hasil dari kedua kasifikasi tersebut manakah yang memiliki nilai akurasi tertinggi. Dari hasil tersebut nantinya akan dipilih dan digunakan untuk melakukan deteksi web phising.

1.6 Sistematika Penulisan

Sistematik penulisan disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Pada penelitian ini terdiri dari beberapa bab yang masing-masing bab mempunyai uraian pokok permasalahan. Untuk uraian tiap bab sebagi berikut :

BAB I PENDAHULUAN

Bab ini akan menuraikan secara ringkas pembahasan tentang latar belakang, rumusan masalah, batasan masalah, maksud dan tujuan, metodologi penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini menjelaskan mengenai teori-teori yang mendukung dalam proses penyusunan penelitian ini. Disini berisi mengenai definisi-definisi

dan teori yang menjadi dasar dalam melakukan penelitian ini. Sumber dari landasan teori ini diambil dari penelitian terkait, buku, jurnal, maupun dari internet.

BAB III ANALISIS DAN PERANCANGAN

Bab ini menjelasakan analisis yang dilakukan peneliti dan menjelaskan langkah-langkah penelitian beserta metode yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelasakan mengenai hasil uji coba terhadap metode yang diimplementasikan. Selain itu pada bab ini juga menjelaskan mengeni analisis hasil uji coba tersebut.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran-saran yang diperoleh dari penelitian yang telah dilakukan.

DAFTAR PUSTAKA

Pada daftar pustaka berisi tentang sumber-sumber yang peneliti gunakan untuk melakukan penelitian.