BAB I
PENDAHULUAN

1.1 Latar Belakang

Phishing merupakan upaya ilegal untuk mencuri informasi penting seperti user ID dan kata sandi melalui situs palsu [1]. Laporan APWG mencatat 165.772 situs phishing terdeteksi pada kuartal pertama 2020, dengan rekor 245.771 serangan di Januari 2021, terutama menargetkan lembaga keuangan [2], [3]. Upaya pencegahan dilakukan, termasuk pengembangan model *machine learning* untuk mendeteksi *phishing*. Dalam mengembangkan *machine learning* berbagai algoritma digunakan sebagai model, salah satunya adalah *Random Forest* [4], [5], [6] yang memiliki akurasi tinggi dalam mendeteksi situs *phishing*, sering kali lebih baik daripada algoritma lain, terutama dalam mengurangi tingkat kesalahan [7].

Algoritma *Random Forest* dikenal karena kemampuannya membangun beberapa pohon keputusan dari berbagai subset data, kemudian menggabungkan hasil prediksi untuk mendapatkan klasifikasi yang lebih akurat [8]. *Random Forest* juga unggul dalam mencegah *overfitting* serta mampu menangani dataset yang kompleks dan tidak seimbang, sehingga menjadi pilihan yang ideal dalam berbagai aplikasi *machine learning*, termasuk deteksi phishing [9]. Namun *random forest* memiliki kompleksitas komputasi yang tinggi, terutama ketika menggunakan banyak pohon keputusan, yang dapat menyebabkan waktu pelatihan dan prediksi menjadi lebih lama [10]. Karena *random forest* memiliki kompleksitas komputasi tinggi maka digunakan seleksi fitur sebagai solusi [11]. Pemilihan fitur efektif mengurangi kompleksitas model dengan mengidentifikasi dan mempertahankan fitur yang penting, sehingga meminimalkan informasi yang tidak relevan dan meningkatkan kinerja pengklasifikasi [12].

Berhubungan dengan pemilihan fitur para peneliti telah menggunakan kombinasi *Random Forest, Decision Tree,* dan *Naïve Bayes* dengan *Pearson Correlation* untuk memilih fitur yang digunakan, dimana metode ini efektif dalam mengurangi jumlah fitur berdasarkan nilai tinggi dari *pearson correlation* [4].

Namun, metode ini bergantung pada hubungan linear, yang membuatnya lebih cocok untuk dataset numerik. *Recursive Feature Elimination (RFE)* dalam mendeteksi *spyware* dapat mengoptimalkan kinerja model dengan fitur-fitur paling relevan, sekaligus mengurangi kompleksitas model yang diterapkan dalam *Decision Tree* [13]. Hasil seleksi fitur *RFE* bergantung pada performa model yang digunakan [14]. Pada dataset nominal, *RFE* cenderung unggul karena pemilihannya tidak bergantung pada hubungan linear antar fitur [15]. Selain *RFE*, *chi-square* metode yang efektif dalam seleksi fitur karena membantu mengurangi waktu dan kompleksitas komputasi [16]. *Chi-square* mengevaluasi hubungan antar variabel kategoris dengan membandingkan frekuensi yang diamati dan diharapkan, sehingga cocok untuk dataset nominal [17].

Dalam *Random Forest*, parameter seperti jumlah pohon atau kedalaman maksimum pohon dapat mempengaruhi hasil model yang dihasilkan [18]. Pengaturan parameter pada *random forest* yang dilakukan dengan *Grid Search* maupun *Random Search* berhasil meningkatkan kinerja model dan menghindari *overfitting* [18]. *Random Search* kurang efektif dalam menemukan konfigurasi yang optimal, sedangkan *Grid Search* lebih lama dalam komputasi [19]. Penerapan *Bayesian Optimization* pada algoritma *Random Forest* menghasilkan akurasi kisaran sama dibanding *Grid Search* [20], namun membutuhkan waktu komputasi yang jauh lebih cepat dibanding *Grid Search* [21]. Sehingga dalam penelitian ini, pemilihan fitur dilakukan menggunakan metode *Chi-Square* dan *RFE*, sedangkan *hyperparameter tuning* diterapkan menggunakan *Bayesian Optimization*.

1.2 Rumusan Masalah

Berdasar latar belakang yang telah diuraikan diatas, maka rumusan masalah dalam penelitian ini sebagai berikut:

1. Bagaimana pengaruh penyesuaian *hyperparameter* pada algoritma *Random Forest* dalam meningkatkan akurasi deteksi situs web *phishing* pada dataset *Phishing Website*?
2. Bagaimana pengaruh teknik seleksi fitur dalam meningkatkan kinerja algoritma *Random Forest* dalam mendeteksi situs web *phishing*?

1.3 Batasan Masalah

Dalam melakukan penelitian ini, penulis menetapkan batasan ruang lingkup untuk mencegah kemungkinan kesalahan dalam pelaksanaan penelitian. Parameter masalah yang ditegaskan dalam proses penelitian ini mencakup:

1. Dataset yang digunakan berupa dataset yang diambil dari *UCI Machine Learning Repository*. Dataset ini dikumpulkan dari berbagai sumber, termasuk *Phishtank, MillerSmiles archive*, dan *Google search operators*.
2. Peneliti berfokus pada pengembangan model menggunakan algoritma *Random Forest*.
3. *Penelitian ini hanya akan berfokus pada deteksi situs web phishing.*
4. Penelitian ini berfokus pada parameter – parameter *random forest* yaitu: jumlah pohon, kedalaman maksimum pohon, minimum sampel, penggunaan *bootstrap sampling*, jumlah maksimum fitur.
5. Penelitian ini hanya akan menggunakan teknik seleksi fitur *RFE* dan *Chi-Square*

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah penelitian ini bertujuan untuk :

1. Mengetahui pengaruh penyesuaian *hyperparameter* pada *algoritma Random Forest* dalam meningkatkan akurasi deteksi situs web *phishing* pada dataset *Phishing Website*.
2. Mengetahui pengaruh teknik seleksi fitur dalam meningkatkan kinerja algoritma *Random Forest* dalam mendeteksi situs web *phishing*.

1.5 Manfaat Penelitian

Penelitian ini menawarkan berbagai manfaat yang bisa dimanfaatkan baik dalam konteks teoritis maupun praktis, di antaranya:

1. Manfaat secara teori

Penelitian ini dapat dijadikan referensi untuk memperluas pemahaman mengenai faktor-faktor yang mempengaruhi kinerja model machine learning, seperti pemilihan fitur yang tepat menggunakan *feature*

selection, penyetelan *hyperparameter tuning* yang optimal, serta penerapan algoritma *Random Forest* dalam proses klasifikasi *web phishing*. Melalui pendekatan ini, penelitian memberikan wawasan yang lebih dalam tentang bagaimana pemilihan fitur dan pengaturan hyperparameter berkontribusi pada peningkatan akurasi dan efisiensi model, khususnya dalam mendeteksi situs phishing yang berbahaya.

2. Manfaat secara praktisi

Bagi lembaga keamanan siber lainnya, sistem klasifikasi berbasis *machine learning* ini dapat digunakan untuk meningkatkan deteksi dini serangan phishing, sehingga melindungi data dan informasi publik serta sektor-sektor kritis. Model ini juga dapat membantu dalam pengawasan dan regulasi internet, memblokir situs phishing secara otomatis, serta memberikan perlindungan yang lebih efektif bagi masyarakat. Selain itu, sistem ini dapat mendukung lembaga keuangan dan bisnis dalam mengidentifikasi situs phishing yang menargetkan pelanggan mereka, sehingga meningkatkan keamanan transaksi online di Indonesia.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN, Bab ini menguraikan latar belakang, perumusan masalah, batasan penelitian, tujuan, manfaat penelitian, serta struktur penulisan.

BAB II TINJAUAN PUSTAKA, Bab ini memuat sejumlah jurnal dan teori dasar yang menjadi landasan penelitian ini.

BAB III METODE PENELITIAN, menjelaskan alat dan bahan yang digunakan dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN, Bab ini membahas tahapan penelitian, termasuk pembahasan kode, pemilihan fitur dataset, implementasi algoritma, peningkatan parameter, dan hasil evaluasi kinerja model.

BAB V PENUTUP, Berisi kesimpulan dan saran dari penelitian yang dilakukan.