BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Dibalik perkembangan teknologi informasi dan internet yang cepat ini, terdapat ancaman-ancaman yang dapat merugikan pengguna teknologi dan internet ini yang biasa disebut *cybercrime* [1]. Salah satu *cybercrime* yang sering terjadi pada saat ini yaitu *phishing*. Serangan *phishing* merupakan *cybercrime* yang menggunakan rekayasa sosial untuk menipu pengguna dan mencuri data informasi para korban, seperti identitas pribadi, informasi penting yang berkaitan dengan keuangan, dan lain-lain. Serangan *phishing* dapat dilakukan dengan beberapa cara, seperti mengirimkan pesan palsu melalui *email* atau platform media sosial, dan juga *website*. Pengguna yang tidak waspada terhadap serangan ini biasanya akan diminta untuk memasukkan informasi atau mengunduh file yang berisi *malware* yang dapat mencuri data pribadi *device* korbannya [2].

Website phishing sendiri merupakan salah satu masalah utama dalam keamanan *website*. Pelaku *website phishing* biasanya mengirimkan *Uniform Resource Locator (URL)* kepada para korban lewat email, sms, dan sosial media [3]. Pada kuartal kedua tahun 2023, *Anti-Phishing Working Group* (APWG) mencatat bahwa terdapat 1.286.208 serangan *phishing* dan terdapat 597.789 serangan *website phishing* yang terdeteksi [4]. Kerugian yang dihasilkan dari serangan ini cukup besar, sehingga perlu dilakukan perbaikan.

Website phishing merupakan replika dari suatu situs *website* yang resmi. Seluruh *website phishing* tidak dibangun untuk *phishing*, tetapi hanya beberapa halaman yang memang dikhususkan oleh pelaku untuk memberikan *input* atau *download*, sehingga ketika data terkirim, data tersebut dikirim ke penyerang. Beberapa metode dalam melakukan kloning *website*, yaitu bisa dengan menggunakan *software* khusus atau bisa dengan membuat secara manual [5]. Dalam mendeteksi *website phishing*, terdapat fitur-fitur yang dapat dijadikan data untuk analisis dan dibagi dalam beberapa format, yaitu *URL-based-feature* yang merupakan fitur yang didapatkan dengan menganalisis teks dari URL secara sederhana. Kemudian untuk fitur lainnya yaitu *content-based feature* yang memiliki fungsi untuk melakukan ekstraksi dengan memuat halaman web dari URL dan menganalisis konten HTML web tersebut. Konten HTML web pada fitur ini,

seperti konten *hyperlink* dan konten *abnormal*, Kemudian untuk fitur terakhir yaitu, *external service feature* yang memiliki fungsi sebagai fitur yang didapatkan dengan melakukan *query* pada layanan pihak ketiga referensi dan mesin pencari, seperti WHOIS, Alexa, Openpagerank, dan Google [1].

Support Vector Machine (SVM) merupakan salah satu algoritma pada *machine learning*. Dalam SVM, setiap *item* data dipetakan sebagai titik dalam ruang n-dimensi dan algoritma ini membangun garis pemisah untuk klasifikasi dua kelas yang dikenal dengan *hyperplane* [6]. Kelebihan dari algoritma SVM adalah dapat mengelompokan data dengan jumlah kecil dan dalam waktu operasi yang tidak lama [7].

Penelitian yang dilakukan oleh Dogukan Aksu dkk. [8] untuk mendeteksi *website phishing* dengan menggunakan algoritma SVM. Penelitian tersebut menggunakan enam atribut dari URL, yaitu *Long URL*, *Dots, IP Address, SSL Connection, At (@) Symbol*, dan *Dash(-) symbol* yang digunakan sebagai fitur utama dalam pelatihan model SVM. Pada penelitian tersebut algoritma SVM dapat melakukan deteksi *website phishing* dengan fitur-fitur ekstraksi yang digunakan dari dataset pelatihan untuk klasifikasi URL. Penelitian tersebut, mendapatkan hasil akurasi yaitu 95%, kemudian nilai recallnya yaitu 88%, nilai precision yaitu 91,66%, dan F1 score yaitu 89.79%. Namun, fitur-fitur yang di ekstraksi memiliki kekurangan yaitu jika terdapat *URL* yang diuji dan memiliki fitur yang sebelumnya pada saat pelatihan tidak pernah diekstaksi sebelumnya, maka sistem akan memberikan hasil yang salah. Selain itu, jika fitur-fitur yang diesktraksi diubah, maka dapat mempengaruhi hasil, sehingga perlu menentukan fitur secara dinamis [8].

Pada penelitian yang dilakukan oleh Muhammad Hasan dkk. pada tahun 2021 [9], deteksi *website phising* menggunakan enam algoritma, yaitu *Logistic Regression, KKN, Decission Tree, Random Forest, SVM*, dan *Gradient Boosting*. Pada penelitian tersebut algoritma SVM memiliki tingkat performa yang paling rendah karena kekurangan yang dimiliki oleh SVM, yaitu tidak dapat melakukan pelatihan model dengan dataset yang cukup besar dan fitur yang banyak. Penggunaan linear kernel juga tidak dapat membantu pada data yang *noisy* dan *target class* yang *overlap*, Sehingga dapat disimpulkan bahwa algoritma SVM tidak dapat digunakan untuk dataset dalam jumlah yang besar dan fitur yang banyak [9].

Pada penelitian kali ini, penelitian dilakukan dengan menggunakan algoritma *Support vector Machine (SVM)* sebagai algoritma utama untuk melakukan klasifikasi. Kemudian pada penelitian yang dilakukan Dogukan Aksu

[8], penelitian yang dilakukan masih memiliki kekurangan dan pada penelitian Hasan pada tahun 2021 juga memiliki banyak kendala seperti data yang cukup besar dan fitur yang banyak dan keduanya menggunakan algoritma SVM. Dari penelitian-penelitian yang menggunakan algoritma SVM, penelitian kali ini akan menggunakan algoritma yang sama, namun akan ditambahkan dengan model seleksi fitur. Fitur seleksi yang digunakan yaitu *Recursive Feature Elimination* yang memiliki keunggulan dalam melakukan filter dataset dan dapat melakukan optimasi pada akurasi suatu model *Machine Learning*.

Recursive Feature Elimination (RFE) merupakan salah satu metode *feature selection* yang melakukan pemilihan model berdasarkan model yang dipelajari dan akurasi klasifikasi. RFE secara berurutan menghilangkan fitur yang tidak penting yang dapat menyebabkan penurunan akurasi klasifikasi sehingga setelah mendapatkan fitur yang terbaik, teknik ini membangun kembali model klasifikasi yang baru. Model ini melakukan pelatihan dengan dataset pelatihan, bobot fitur yang mencerminkan pentingnya setiap fitur yang diperoleh. Fitur-fitur yang sudah diurutkan berdasarkan bobot tertinggi, akan diklasifikasikan ulang sehingga penggunaan metode RFE berbasis kepentingan fitur dapat diperoleh [10].

Pada penelitian yang dilakukan oleh Puneet Misra dan Arun Singh Yadav pada tahun 2020 [11] mengenai optimasi klasifikasi pasien diabetes menggunakan *Recursive Feature Elimination*, RFE mempunyai peranan penting dalam meningkatkan akurasi model klasifikasi SVM. Dalam penelitian tersebut, penggunaan model RFE sangat membantu dalam meningkatkan performa algoritma SVM. Hasil dari penelitian pun cukup baik karena model RFE dapat melakukan *filter* data validasi silang hingga 10 kali lipat setelah memasukkan fitur teratas. Hal ini dapat menghindari *overfitting* sehingga proses *filter* mendapatkan hasil terbaik. RFE juga membantu model algoritma SVM untuk mendapatkan tingkat akurasi yang lebih tinggi dibandingkan dengan model algoritma tanpa RFE. Oleh karena itu, penelitian ini dilakukan untuk menguji apakah fitur seleksi menggunakan RFE dapat membantu algoritma SVM dalam mendapatkan tingkat akurasi yang lebih tinggi atau tidak.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang sudah dijelaskan sebelumnya, terdapat rumusan masalah dalam penelitian ini, yaitu :

1. Bagaimana mengimplementasikan seleksi fitur *Recursive Feature*

Elimination yang digabungkan dengan algoritma *Support Vectore Machine* dalam mendeteksi *website phishing*?

2. Berapa nilai *precision*, *recall*, F1 *score*, dan akurasi pada deteksi *website phishing* dengan seleksi fitur RFE dan algoritma SVM?

1.3 Batasan Permasalahan

Agar tidak terjadi penyimpangan dari judul dan tujuan yang ingin dicapai, maka terdapat batasan masalah pada penelitian ini, yaitu :

1. Dataset diambil dari Mendeley Data dengan jumlah data yaitu 11.430 data dengan 87 atribut
2. Data untuk *training* sebanyak 70% dari data dan data untuk *test* sebanyak 30% dari data.

1.4 Tujuan Penelitian

Dari proses penelitian ini, terdapat tujuan yang ingin dicapai, yaitu :

1. Mengimplementasikan seleksi fitur *Recursive Feature Elimination* yang digabungkan dengan algoritma *Support Vectore Machine* dalam mendeteksi *website phishing*
2. Mengukur nilai *precision*, *recall*, F1 *score*, dan akurasi dari fitur seleksi RFE dan algoritma SVM dalam deteksi *website phishing*

1.5 Manfaat Penelitian

Proses penelitian yang dilakukan ini memiliki manfaat untuk memberikan informasi mengenai kinerja model *Recursive Feature Elimination* dalam seleksi fitur untuk optimasi model algoritma SVM dalam meningkatkan akurasi deteksi *website phishing*.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN

  Bab I bagian pendahuluan berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan juga sistematika penulisan dari skripsi yang berjudul ”IMPLEMENTASI RECURSIVE FEATURE ELIMINATION DAN SUPPORT VECTOR MACHINE UNTUK DETEKSI WEBSITE PHISHING.”

- Bab 2 LANDASAN TEORI

  Bab II bagian landasan teori berisi tentang teori-teori dan metode-metode yang berkaitan dengan penelitian yang dilakukan. Teori dan metode yang digunakan, yaitu *website phising*, *Recursive Feature Elimination* (RFE), *Support Vector Machine* (SVM), *Confusion Matrix Binary*.

- Bab 3 METODOLOGI PENELITIAN

  Bab III bagian metodologi penelitian berisi tentang penjelasan mengenai metode-metode yang digunakan untuk penelitian ini. Isi dari Bab tiga ini berisi *flowchart* utama yang berisi proses penelitian, *flowchart preprocessing, flowchart RFE*, dan *flowchart SVM*

- Bab 4 HASIL DAN DISKUSI

  Bab empat bagian hasil dan diskusi berisi tentang hasil implementasi yang dilakukan menggunakan metode-metode yang dilakukan dan hasil uji coba dari algoritma *support vector machine* dengan *feature selection Recursive Feature Elimination* menggunakan dataset yang sudah disiapkan sebelumnya.

- Bab 5 KESIMPULAN DAN SARAN

  Bab lima bagian kesimpulan dan saran berisi tentang kesimpulan dari hasil uji coba yang dilakukan pada penelitian ini dan saran untuk pengembangan lebih lanjut di masa depan mengenai topik penelitian ini.