BAB I
PENDAHULUAN

1.1 Latar Belakang

Kemajuan teknologi mempercepat penyebaran informasi. Saat ini, jutaan perangkat dan penggunanya terhubung melalui Internet, aksesnya yang user friendly membuat seseorang dapat dengan mudah menerima sebuah informasi melalui internet. Email adalah cara yang efektif, sederhana, cepat, dan murah untuk berkomunikasi. Secara umum konsep penggunaan email sama seperti surat biasa, yang membedakan adalah email di dukung internet maka email memiliki beberapa keunggulan seperti dapat mengirimkan lebih dari 1 orang, waktu pengiriman cepat, hemat, dan dapat memuat informasi dalam bentuk file dokumen teks atau spreadsheet, gambar, audio dan video [1]. Lalu lintas email bisa mencapai 319 miliar per hari pada tahun 2021 [2]. Jumlah email yang sangat besar inilah yang membuat banyak orang menyalahgunakan email sebagai ajang promosi iklan, phishing, bahkan mengirimkan virus. Email yang tidak penting bagi pengguna email dapat disebut dengan email SPAM *(Stupid Pointless Annoying Message)*. Metode ini dapat dikirim dalam volume besar oleh bot atau botnet, serta jaringan komputer yang terinfeksi.

Berdasarkan penelitian keamanan siber senior di Kaspersky pada tahun 2021, mengungkap bahwa lebih dari 267 miliar diantaranya (83,69 persen) merupakan email SPAM yang dikirim dan diterima. Artinya Cuma 52 miliar (16,3 persen) yang merupakan email sungguhan. SPAM lebih dari sekedar mengganggu, dia bisa berbahaya terutama jika itu adalah bagian dari penipuan dengan metode phising untuk mendapatkan kata sandi, nomor kartu kredit, rincian bank, dan lainnya. Pada umumnya email spam menunggangi executable, MS Office, text, PDF, JS, RAR, ISO. Kawasan Asia Pasifik mendapatkan porsi besar spam email 24% persen per 2022. Untuk Indonesia sendiri menduduki peringkat ke 4 dengan jumlah 10,4 % [2].

Salah satu cara untuk meminimalisir masalah ini yaitu dengan melakukan penyaringan email SPAM yang berdasarkan kandungan dari email itu sendiri. Dari

beberapa penelitian sebelumnya, untuk melakukan penyaringan email SPAM digunakan seleksi fitur dan algoritma klasifikasi. Algoritma yang digunakan diantaranya adalah random forest dan logistic regression. Pada penelitian [3] dilakukan perbandingan algoritma Naïve Bayes, SVM, J48 ,dan random forest untuk mengidentifikasi email SPAM. Dari hasil pengujian, random forest memiliki nilai akurasi paling tinggi. Sedangkan pada penelitian [4] nilai akurasi logistic regression lebih tinggi dibanding random forest.

Berdasarkan permasalahan yang ada, maka peneliti memutuskan menggunakan algoritma random forest dan logistic regression untuk mengidentifikasi email SPAM menggunakan seleksi fitur information gain. Metode random forest digunakan pada penelitian ini karena metode ini mampu menangani input variabel yang besar dan dapat menyeimbangkan error dalam unbalanced dataset [5]. Dan logistic regression diformulasikan klasifikasi data ke dalam dua group dan menjelaskan variabel biner, maka logistic regression cocok digunakan untuk memprediksi keanggotaan variabel independen dalam dua grup saja [6]. Sedangkan information gain dapat di gunakan untuk menghilangkan noise pada fitur-fitur yang tidak relevan [7]. Maka dengan demikian peneliti memutuskan memilih judul "Perbandingan Algoritma Random Forest dan Logistic Regression dengan Seleksi Fitur Information Gain untuk Klasifikasi Email SPAM".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, maka rumusan masalah pada penelitian ini adalah "Bagaimana hasil perbandingan antara algoritma *Random Forest* dan *Logistic Regression* sebelum dan setelah dilakukan seleksi fitur *Information Gain* untuk klasifikasi email SPAM?".

1.3 Batasan Masalah

Batasan masalah pada penelitian ini sebagai berikut

1. Penelitian ini hanya membandingkan dua algoritma yaitu *Random Forest* dan *Logistic Regression*.

2. Penelitian ini dilakukan untuk mengetahui nilai akurasi pada algoritma *Random Forest* dan *Logistic Regression* sebelum dan sesudah dilakukan seleksi fitur.
3. Penelitian ini dilakukan untuk mengetahui fitur-fitur apa saja yang relevan terhadap hasil klasifikasi.
4. Penelitian ini dilakukan untuk mengetahui apakah ada peningkatan atau penurunan kinerja *Random Forest* dan *Logistic Regression* setelah dilakukan seleksi fitur
5. Dataset yang digunakan yaitu data email SPAM dan ham yang didapat dari UCI *Machine Learning Repository*.
6. Dalam penelitian ini menggunakan seleksi fitur *Information Gain*.
7. Pada penelitian ini menggunakan google colab untuk melakukan klasifikasi.

1.4 Tujuan Penelitian

Tujuan penelitian ini yaitu untuk mengetahui hasil perbandingan antara algoritma *Random Forest* dan *Logistic Regression*, serta mengetahui pengaruh seleksi fitur terhadap algoritma klasifikasi *Random Forest* dan *Logistic Regression*.

1.5 Manfaat Penelitian

Manfaat dari penelitian yang dilakukan adalah sebagai berikut :

1. Bagi Pembaca
   a. Sumber informasi tentang pengklasifikasian email SPAM.
   b. Sebagai bahan pertimbangan untuk peneliti selanjutnya.
2. Bagi Penulis

   Hasil penelitian ini oleh peneliti diharapkan dapat bermanfaat untuk:
   a. Menambah wawasan penulis mengenai Email SPAM.
   b. Mengetahui algoritma klasifikasi yang digunakan untuk mengidentifikasi email SPAM.

1.6 Metode Penelitian

Tahapan penelitian ini dimulai dengan studi literatur dari penelitian terkait. Kemudian mengambil dataset dari UCI *Machine Learning Repository*. Selanjutnya seleksi fitur diterapkan pada dataset untuk menyeleksi fitur memiliki tingkat relevansi tinggi satu sama lain. Fitur yang memiliki tingkat relevansi yang rendah tidak akan digunakan. Kemudian hasil dari seleksi tersebut nantinya akan digunakan untuk menguji metode klasifikasi *Random Forest* dan *Logistic Regression*. Sehingga dari hasil tersebut bisa dibandingkan manakah tingkat akurasi yang paling tinggi.

1.7 Sistematika Penulisan

Sistematika penulisan skripsi ini dibagi dalam beberapa bab pokok permasalahan sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian dan sistematika penulisan penelitian.

BAB II LANDASAN TEORI

Bab ini di jelaskan mengenai landasan teori dan kajian pustaka. Sumber dari landasan teori ini diambil dari penelitian terkait, buku, jurnal, maupun dari internet.

BAB III METODE PENELITIAN

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

BAB IV HASIL DAN PEMAHASAN

Pada 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 merupakan akhir dari skripsi yang berisi kesimpulan, dan juga saran dari penelitian ini.