BAB I

PENDAHULUAN

1.1 Latar Belakang

Email atau surat elektronik adalah salah satu metode komunikasi yang paling umum dan banyak dipakai di era digital. Sejak penemuannya pada tahun 1971, email telah melalui berbagai evolusi dan peningkatan fungsi untuk memenuhi permintaan pengguna yang terus berkembang. Pada artikel *The History Of Spam* oleh Justin A. Smith, Spam email dimulai pada tahun 1978 dengan pengiriman pesan iklan pertama oleh Gary Thuerk selama 1990-an. Email spam adalah pesan email yang tidak diminta atau tidak diinginkan yang dikirim secara massal ke ribuan atau jutaan alamat email, email spam biasanya berisi iklan atau promosi produk, upaya phishing, virus atau malware, dan pesan palsu lainnya. Sedangkan email ham merujuk pada email yang dianggap tidak spam, yaitu pesan yang diinginkan bagi penerima. Ini mencakup komunikasi dari keluarga, teman atau rekan kerja, serta *newsletter* dan informasi penting. [1]

Seiring berkembangnya internet, email spam meningkat karena dimanfaatkan oleh berbagai pihak, termasuk perusahaan, untuk pemasaran massal dengan biaya rendah. Selain iklan produk, spam sering digunakan dalam skema penipuan, contohnya scam "*Nigerian Prince*" yang menjanjikan imbalan besar dengan alasan bantuan transfer uang. Penipuan lainnya berupa notifikasi pengiriman paket palsu, terutama saat musim belanja, guna memancing korban membuka tautan berbahaya. Modus-modus seperti ini menimbulkan risiko signifikan, seperti pencurian identitas dan kerugian finansial.

Dengan adanya masalah seperti di atas metode yang efektif untuk mengatasi masalah ini sangat penting, karena dapat memastikan solusi yang tepat sasaran, efisien, dan berkelanjutan, sehingga tidak hanya memperbaiki kondisi saat ini tetapi juga mencegah masalah serupa muncul di masa depan. Klasifikasi email spam dan ham merupakan solusi yang umum digunakan untuk mengatasi masalah seperti email spam. Terdapat banyak algoritma *machine learning* yang bisa digunakan untuk menyelesaikan masalah ini, tetapi performa masing-masing

algoritma bisa berbeda tergantung pada jenis data dan metode evaluasi yang digunakan. Di antaranya menggunakan algoritma *machine learning* yaitu algoritma Decision Tree, algoritma ini adalah algoritma *machine learning* yang digunakan untuk klasifikasi dan regresi. Dan memiliki versi algoritma seperti CART, C4.5, dan C5.0.

CART dikenal sebagai salah satu metode pembelajaran mesin berbasis pohon keputusan yang efektif dalam pengklasifikasian dan analisis regresi. Kelebihan dari algoritma CART adalah kemampuannya untuk menghasilkan model yang mudah diinterpretasikan dan dipahami, karena representasi pohon keputusan yang jelas. Selain itu, CART dapat menangani data dengan variabel numerik dan kategorikal, serta mampu mengatasi masalah interaksi antar variabel. Namun, kekurangan CART meliputi kecenderungannya untuk overfitting, terutama pada dataset kecil.[2]

C4.5 merupakan pengembangan evolusi dari Iterative Dichotomiser 3 (ID3) yang diciptakan oleh J. Ross Quinlan. Algoritma ini diterapkan untuk mengklasifikasikan data yang mengandung elemen numerik dan kategorikal. Kelebihan dari C4.5 adalah untuk menangani dataset yang memiliki variabel numerik dan kategorikal, serta menghasilkan pohon keputusan yang lebih seimbang dan kurang rentan terhadap overfitting dibandingkan dengan algoritma sebelumnya, seperti ID3. Namun, kekurangan dari C4.5 adalah waktu komputasi yang lebih lama, terutama pada dataset yang besar dan kompleks, serta ketergantungannya pada ukuran dataset yang dapat mempengaruhi akurasi jika data tidak cukup representatif.[2]

C5.0 merupakan salah satu metode dalam data mining yang secara khusus dirancang untuk membuat pohon keputusan dalam proses klasifikasi. Algoritma C5.0 ini merupakan perkembangan dari dua algoritma sebelumnya, yaitu ID3 dan C4.5. Algoritma C5.0 menyediakan peningkatan efisiensi dan kecepatan dalam pemrosesan data, menangani variabel yang hilang dan menghasilkan model yang lebih ringkas dengan akurasi yang lebih tinggi dibandingkan dengan C4.5. Namun, kekurangan dari C5.0 adalah biaya lisensi untuk penggunaan perangkat lunak komersialnya, rentan terhadap noise dalam data dan dapat menghasilkan

model yang kurang intuitif jika struktur pohon terlalu kompleks.[3]

Beberapa penelitian terdahulu telah membahas klasifikasi email spam, tidak hanya berfokus pada objek email, tetapi juga menggunakan objek lain. Penelitian-penelitian tersebut juga menerapkan algoritma seperti CART, C4.5, dan C5.0 untuk klasifikasi. Fida Maisa Hana et al (2023) menggunakan algoritma CART dan berhasil mencapai akurasi sebesar 100% tanpa pemangkasan dan pemangkasan awal. Sementara itu, penggunaan pemangkasan dan pemangkasan awal menghasilkan tingkat akurasi sebesar 96,15% dalam klasifikasi penyakit diabetes. Sementara itu, studi lain oleh Agung Purwanto dan Handoyo Widi Nugroho (2023) dalam penelitian analisa perbandingan kinerja algoritma C4.5 dan algoritma k-nearest neighbors untuk klasifikasi penerima beasiswa dan menemukan bahwa C4.5 lebih rendah dengan akurasi 97,23%, Rahmanita Widiyanti et al (2022) Untuk penelitian Implementasi Algoritma C5.0 Untuk Klasifikasi Kepuasan Masyarakat Terhadap Pelayanan Kantor menghasilkan akurasi sebesar 90,67% pada dataset kecil. [4],[5],[6]

Berdasarkan penjabaran permasalahan diatas, maka akan dilakukan Penelitian “Perbandingan Kinerja Algoritma CART (Classification and Regression Tree), C4.5, dan C5.0 Dalam Klasifikasi Email Spam dan Ham (non-spam). Dengan pendekatan kuantitatif dan mengolah data menggunakan *tools Rapidminer*. Dataset yang digunakan berasal dari dataset text email spam dan ham dengan dua atribut yang tersedia di repository Kaggle. Dengan data yang sudah terlabel ini memudahkan proses perbandingan algoritma dalam klasifikasi email spam dan ham, karena data sudah melalui tahap pengumpulan dan pengolahan awal meskipun harus dilakukan proses text preprocessing data dan pembobotan fitur menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*).

1.2 Rumusan Masalah

Berdasarkan permasalahan yang teridentifikasi, rumusan masalah dalam penelitian ini yaitu Bagaimana perbandingan nilai akurasi algoritma CART, C4.5, dan C5.0 dalam mengklasifikasikan email sebagai spam dan ham (non-spam)?

1.3 Ruang Lingkup Masalah

Adapun ruang lingkup masalah yang akan dibahas dalam penelitian ini agar penelitian yang dilakukan bisa lebih terarah dan tidak keluar dari pokok utama permasalahan yaitu :

1. Dataset yang digunakan adalah data public yaitu data email spam yang tersedia di situs *Kaggle* dengan jumlah data 2.638 data.
2. Penelitian ini akan fokus pada tiga algoritma yaitu Algoritma CART (*Classification and Regression Trees*), C4.5, dan C5.0. Masing-masing algoritma akan dibandingkan hasil akurasi dalam konteks klasifikasi email spam dan ham.
3. Tools/Aplikasi pendukung yang digunakan dalam perhitungan penelitian ini adalah *Tools Rapidminer*.

1.4 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah maka tujuan dari penelitian ini untuk membandingkan hasil akurasi kinerja algoritma CART, C4.5, dan C5.0 dalam klasifikasi email sebagai spam dan ham (non-spam), dengan menggunakan dataset email spam yang tersedia di *Kaggle* dan *Tools RapidMiner*, serta memberikan rekomendasi algoritma yang paling efektif untuk sistem klasifikasi email.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut :

1. Penelitian ini dapat membantu dalam mengklasifikasikan email sebagai spam atau ham secara lebih efektif, sehingga meningkatkan keamanan pengguna dari potensi penipuan dan pencurian identitas yang sering terjadi melalui email spam.
2. Dengan memahami kinerja masing-masing algoritma, pengguna dan pengembang sistem klasifikasi email mampu mengambil keputusan yang lebih tepat dalam memilih algoritma yang paling sesuai dengan keperluan

mereka, baik dari segi akurasi dan hasil perhitungan lainnya.

3. Meningkatkan literatur dan penelitian lanjutan tentang penerapan *machine learning* untuk keamanan digital.

1.6 Sistematika Penulisan

Adapun sistematika penulisan penelitian yang dibutuhkan sebagai berikut :

BAB I : PENDAHULUAN

Pada bab ini menguraikan tentang latar belakang masalah, ruang lingkup, perumasan masalah, tujuan penelitian, serta sistematika penulisan.

BAB II : TINJAUAN PUSTAKA

Pada bab ini menjelaskan landasan teori yang terkait dengan masalah yang diteliti, kerangka pemikiran serta hipotesis penelitian.

BAB III : METODE PENELITIAN

Pada bab ini menguraikan tentang metode penelitian yang terdiri dari metode pengumpulan data, objek penelitian, alat pendukung, metode penelitian, tahap penelitian dan pengolahan data.

BAB IV : HASIL DAN PEMBAHASAN

Pada bab ini diuraikan mengenai garis besar tentang penelitian dari deskripsi data, hasil, dan implementasi.

BAB V : SIMPULAN DAN SARAN

Pada bab ini uraikan mengenai kesimpulan dari hasil penelitian. Isi dari bab ini meliputi kesimpulan dan saran.

DAFTAR PUSTAKA

LAMPIRAN