

BAB I PENDAHULUAN

1.1 Latar Belakang

Berbagai aspek kehidupan manusia sudah sangat dipengaruhi oleh kemajuan pesat dalam teknologi informasi dan komunikasi, termasuk cara individu dan organisasi bertukar informasi. Kehadiran *email* atau surat elektronik, merupakan inovasi penting dalam hal ini. Karena *email* memungkinkan orang untuk berkomunikasi secara cepat, efektif, dan murah tanpa batasan waktu atau lokasi, *email* telah menjadi alat komunikasi digital yang sangat populer. Namun, seiring dengan peningkatan jumlah pengguna dan kepercayaan yang meningkat terhadap *email* sebagai alat komunikasi utama, muncul masalah baru yang cukup serius yaitu penyebaran *spam email* [1].

Spam email adalah pesan yang dikirim secara massal tanpa diminta oleh penerima dan sering kali berisi iklan, penipuan, atau bahkan tautan yang dapat membahayakan keamanan data pengguna. *Spamming* tidak hanya membuat pengguna tidak nyaman, tetapi juga dapat mengurangi produktivitas dan meningkatkan risiko serangan siber. Oleh karena itu, menghilangkan *spam email* menjadi sangat penting untuk menjaga kualitas komunikasi digital di era *modern*. Pendekatan berbasis pembelajaran mesin merupakan solusi yang baik untuk menghilangkan *spam email* karena dapat melakukan pembelajaran dari data dan mengenali pola baru secara otomatis [2] [3].

Klasifikasi *spam email* dalam *machine learning* termasuk ke dalam masalah klasifikasi *biner*, yaitu mengklasifikasikan *email* sebagai *spam* atau *not-spam* (ham). *Decision Tree*, *Naive Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* adalah beberapa algoritma yang umum digunakan untuk klasifikasi email spam. *Naive Bayes* dikenal sangat cepat dan efisien untuk teks besar, tetapi dia mengasumsikan independensi antar fitur yang seringkali tidak mungkin pada teks alami, sehingga akurasi dapat menurun pada dataset yang kompleks. Meskipun SVM dapat menghasilkan perpecahan kelas yang baik di ruang besar (seperti

representasi TF-IDF), proses pelatihannya agak lambat, dan modelnya sulit diolah oleh manusia. *Random Forest*, yang menawarkan ketahanan terhadap *overfitting* dan stabilitas yang tinggi berkat kombinasi banyak keputusan pohon, membutuhkan lebih banyak sumber daya komputasi dan kurang mudah dipahami dibandingkan dengan model pohon tunggal [6] [8] [9].

Decision Tree memiliki keunggulan, seperti kemudahan interpretasinya, tidak perlu normalisasi data, dan fleksibel dalam menangani data numerik dan kategorikal. Tetapi algoritma ini memiliki masalah dengan *overfitting* yang dapat menyebabkan penurunan kinerja ketika dihadapkan pada data baru [4]. Maka dari itu perlu adanya metode untuk menekan *overfitting* tersebut, salah satunya dengan penerapan *ensemble learning*.

Salah satu pendekatan yang terbukti mampu meningkatkan kinerja klasifikasi adalah *ensemble learning*, yaitu metode yang menggabungkan beberapa model pembelajaran untuk menghasilkan prediksi yang lebih akurat dan stabil [5]. Dua metode *ensemble learning* yang banyak digunakan adalah *Adaptive Boosting* (AdaBoost) dan *Bootstrap Aggregating* (Bagging). AdaBoost bekerja dengan meningkatkan bobot pada data yang salah diklasifikasikan pada iterasi sebelumnya sehingga model menjadi lebih fokus pada data yang sulit diprediksi, sedangkan Bagging membangun beberapa model dari data bootstrap untuk mengurangi *varians* dan menghindari *overfitting* [6][7].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membandingkan kinerja algoritma *Decision Tree* dengan metode *ensemble* Bagging dan AdaBoost dalam klasifikasi *email spam*, serta menentukan metode *ensemble* terbaik yang paling efektif dalam meminimalkan risiko *overfitting* pada *Decision Tree*.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah sebagai berikut :

1. Bagaimana komparasi dari metode Bagging dan Adaboost pada algoritma *Decision Tree* terhadap hasil akurasi pada klasifikasi *email*

spam?

2. Metode manakah yang memberikan performa terbaik dalam mengatasi permasalahan *overfitting*?

1.3 Batasan Masalah

1. Dataset yang digunakan merupakan dataset yang memiliki 2 kelas yaitu *email spam* dan *not-spam* (ham) dengan jumlah 3796 data.
2. Data yang digunakan adalah data yang sifatnya terbuka (open data) yang tersedia di public dari *Kaggle*.
3. Menggunakan metode *ensemble* yaitu Bagging dan AdaBoost.
4. Metode *ensemble* diterapkan pada algoritma *Decision Tree* sebagai *base learner* (pembelajaran dasar).
5. Bahasa pemrograman yang digunakan adalah *python*.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai dalam penelitian diantaranya:

1. Menganalisis dan membandingkan kinerja algoritma *Decision Tree* dengan metode *ensemble* Bagging dan AdaBoost dalam klasifikasi *email spam*.
2. Menentukan metode *ensemble* terbaik yang paling efektif dalam meminimalkan risiko *overfitting*.

1.5 Manfaat Penelitian

Penelitian diharapkan dapat memberikan manfaat sebagai berikut :

1. Secara Teoritis
 - a. Membantu para peneliti di masa depan sebagai bahan referensi
 - b. Menambah pengetahuan dan wawasan ilmiah mengenai penerapan metode *ensemble learning* Bagging dan AdaBoost pada algoritma *Decision tree* untuk tugas klasifikasi *spam email*, termasuk

pemahaman mendalam tentang bagaimana *ensemble* dapat mengatasi masalah *overfitting*.

2. Secara Praktis

Membantu pengembang sistem atau praktisi teknologi informasi dalam memilih metode klasifikasi yang paling efektif untuk klasifikasi *spam email*. Model yang dihasilkan dapat dijadikan dasar pengembangan sistem penyaring *spam email* yang lebih akurat dan adaptif, sehingga mampu meningkatkan kualitas komunikasi digital, mengurangi risiko keamanan, serta meningkatkan produktivitas pengguna *email*.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN, bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, sistematika penulisan.

BAB II TINJAUAN PUSTAKA, bab ini berisi ulasan teoritis yang mendukung pemahaman terhadap pendekatan yang digunakan, serta berisi kajian Pustaka dari penelitian terdahulu yang relevan dengan topik penelitian.

BAB III METODE PENELITIAN, bab ini berisi metode dan analisis yang dilakukan oleh peneliti, serta berisi penjelasan langkah-langkah penelitian.

BAB IV HASIL DAN PEMBAHASAN, bab ini menjelaskan hasil uji terhadap metode yang digunakan, serta menjelaskan analisis uji coba tersebut.

BAB V PENUTUP, bab ini merupakan akhir dari penelitian berisi kesimpulan, kekurangan penelitian dan juga saran yang dapat dikembangkan oleh penelitian berikutnya