

# BAB I PENDAHULUAN

## 1.1 Latar Belakang Masalah

*Spam* atau bisa juga berbentuk *junk mail* adalah penyalahgunaan sistem pesan elektronik (termasuk media penyiaran dan sistem pengiriman digital) untuk mengirim berita iklan dan keperluan lainnya secara massal. Setiap spam yang diterima memakan waktu dan tenaga si penerimanya untuk membaca, menyortir, menghapus, berusaha menolak di kemudian hari. Spam juga bisa memenuhi mailbox, mengakibatkan mailserver sibuk, dan memperlambat layanan lainnya. Walaupun email spam dapat di tangani dari para penyedia layanan internet dari sisi email header terkadang masih ada spam yang lolos dari pengawasan dari sisi konten email.

Untuk melakukan klasifikasi terhadap email spam dapat dilakukan dengan pohon keputusan. Pada tugas akhir ini, akan dibangun sistem yang menerapkan algoritma klasifikasi pohon keputusan C4.5 untuk mengklasifikasikan email spam dan ham (bukan spam). Kemampuan algoritma C4.5 untuk mem-breakdown proses pengambilan keputusan yang kompleks menjadi lebih simple menghasilkan pengambilan keputusan yang lebih menginterpretasikan solusi dari permasalahan.

Decision tree merupakan algoritma learning yang *unstable*, perubahan kecil terhadap training set mengakibatkan perubahan yang besar pada learned *classifier*[1]. Salah satu cara untuk mengatasinya dengan metode ensemble, membentuk beragam *classifier* dengan memanipulasi data training[4]. Untuk itu pada tugas akhir ini algoritma C4.5 menjadi *classifier* dalam metode ensemble. Pendekatan ensemble yang digunakan adalah random forest.

Tugas akhir ini menganalisis performansi random forest dengan algoritma klasifikasi algoritma C4.5 dalam kasus pengklasifikasian terhadap konten dari spam email.

## 1.2 Perumusan Masalah

Dari latar belakang di atas maka masalah-masalah yang dihadapi, yaitu :

1. Bagaimana mengimplementasikan email spam filtering berdasarkan konten dengan menggunakan metode decision tree.
2. Bagaimana cara memanipulasi data training sebelum digunakan pada algoritma klasifikasi.
3. Bagaimana cara mendapatkan model yang terbaik dari metode Random Forest dengan classifier dari Algoritma C4.5.

Batasan dari permasalahan adalah :

1. Data yang digunakan telah melalui tahap *preprocessing*.
2. Data yang diterima sistem dalam bentuk tipe *continuous*.
3. Data inputan berasal dari email dalam bahasa Inggris yang telah di *preprocessing*. Karena dari riset sebagian besar spam email yang menyebar menggunakan bahasa Inggris[6].

### 1.3 Tujuan

Tujuan yang ingin dicapai dalam pembuatan tugas akhir ini adalah sebagai berikut :

1. Mengimplementasi metode decision tree menggunakan algoritma C4.5 pada permasalahan konten email spam.
2. Menerapkan metode ensemble yaitu random forest untuk memaipulasi data yang akan di gunakan pada *classifier*.
3. Menganalisis model yang menerapkan metode random forest dan C4.5 sebagai *classifier* dan menguji menggunakan data testing.

### 1.4 Hipotesis

Pengklasifikasian email spam menggunakan random forest sebagai metode ensemble dengan *classifier*-nya dari algoritma C4.5 menghasilkan klasifikasi email spam yang memiliki akurasi yang baik (minimal 60%).

### 1.5 Metodologi Penyelesaian Masalah

Metode yang digunakan untuk menyelesaikan masalah yaitu :

1. Studi Pustaka  
Bahan bahan dalam penyelesaian masalah didapat dari literatur baik itu jurnal, buku-buku atau informasi lain yang relevan yang berhubungan dengan Algoritma C4.5, metode ensemble, random forest, text categorization, email filtering, dan decision tree.
2. Pengumpulan data  
Data digunakan didapat dari UCI repository dengan dataset bernama Spambase dengan jumlah instan 4601 dan atribut 57.
3. Implementasi sistem  
Pembentukan sistem yang dapat membentuk model yang menerapkan metode Random Forest dengan *classifier* dari algoritma C4.5.

4. Pengujian  
Model yang terbentuk diujikan menggunakan data uji yang telah dipersiapkan untuk melihat kesesuaian hasil prediksi yang dibuat oleh model.
5. Analisis  
Analisis dilakukan terhadap hasil pengujian dari sistem sehingga dapat dibentuk suatu kesimpulan.
6. Perumusan kesimpulan penyelesaian masalah  
Merumuskan kesimpulan terhadap hasil pengujian pada sistem untuk menyelesaikan masalah.