BAB I
PENDAHULUAN

1.1. Latar Belakang

Eletronic mail atau lebih sering kita kenal dengan singkatan *email* merupakan salah satu layanan internet yang paling banyak digunakan. *Email* adalah media komunikasi yang murah, cepat dan mudah penggunaannya. Format *email* terdiri dari sebuah *envelope*, beberapa *field header*, sebuah *blank line* dan *body*. *Email* memiliki sifat data berupa teks yang semi terstruktur dan memiliki dimensi yang tinggi (Rachli, 2007).

Email spam yang dikirim kepada pengguna *email* pada umumnya berisi akan konten-konten merugikan dan berbahaya dinilai merugikan. Adanya hal tersebut sangat mengganggu pengguna *email*, terlebih para perusahaan atau pekerja individual yang sehari-harinya menggunakan sarana *email* sebagai penunjang kerjanya. Pengklasifikasian *email spam* sangat diperlukan dalam menanggulangi hal ini. pengguna *email* tidak perlu risih dengan adanya *spam* tersebut, karena akan diklasifikasi oleh algoritma metode.

Spam adalah praktek pengiriman pesan komersial atau iklan kepada sejumlah besar *newsgroup* atau *email* yang sebetulnya tidak berkeinginan atau tidak tertarik menerima pesan tersebut (Kristiansen, at all. 2005). *spam* adalah pengiriman *email* yang disalahgunakan karena biaya yang murah dan ini dimanfaatkan oleh perusahaan-perusahaan besar untuk mempromosikan suatu produk walaupun si penerima tidak menyetujuinya.

Beberapa penelitian tentang hal tersebut telah dilakukan diantaranya, dalam penelitian berjudul *Content Based Spam Detection in Email using Bayesian Classifier* menyimpulkan bahwa dengan melakukan tiga kali pengujian menemukan nilai akurasi tertinggi mencapai 96% dalam klasifikasi *spam* (Sunil, 2015). Serta penelitian dalam Perbandingan Kinerja Algoritma Id3 Dan C4.5. Hasil pengukuran menunjukkan algoritma ID3 memiliki kinerja yang lebih baik dibandingkan algoritma C4.5.

Metode klasifikasi dibedakan menjadi dua yaitu metode klasifikasi parametrik dan *nonparametrik*. *K-Nearest Neighbor* (KNN)adalah salah satu metode klasifikasi *nonparametrik*. Dalam prosesnya, KNN memeriksa semua kata dalam dokumen pelatihan untuk menghitung kesamaannya dengan dokumen yang akan diklasifikasikan (dokumen uji). KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap *training* data yang memiliki banyak *noise* dan efektif apabila *training* datanya besar. Peneliti menggunakan algoritma KNN karena memiliki konsistensi yang kuat. Ketika jumlah data mendekati tak hingga, algoritma KNN menjamin *error rate* yang minim.

Peneliti memiliki dataset yang bersumber dari *UCI Machine Learning* terhadap klasifikasi *email spam*. Dataset tersebut memiliki perbedaan atribut dari penelitian diatas serta peneliti ingin mencoba metode lain dalam klasifikasi *email spam*. Dari latar belakang tersebut penulis disini memiliki ketertarikan untuk mengangkat sebuah judul yaitu "KLASIFIKASI *E-MAIL SPAM* MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*".

1.2. Rumusan Masalah

Dari latar belakang diatas penulis memilih beberapa pokok permasalahan sebagai berikut:

1. Berapa tingkat akurasi yang diperoleh pada klasifikasi *email spam* menggunakan metode *K-nearest neighbor*?
2. Berapa tingkat presisi yang diperoleh pada klasifikasi *email spam* menggunakan metode *K-nearest neighbor*?

1.3. Tujuan

Dari hasil penelitian ini penulis memiliki beberapa tujuan yaitu:

1. Mengetahui tingkat akurasi tertinggi dari metode *K-Nearest Neghbor* pada klasifikasi *email spam*.
2. Mengetahui tingkat presisi tertinggi dari metode *K-nearest neighbor* pada klasifikasi *email spam*.

1.4. Manfaat

Dalam penelitian ini penulis berharap adanya manfaat terhadap pihak lain atau pengembang nantinya, maka dari itu manfaat hasil penelitian ini yaitu sebagai alternatif metode untuk mengklasifikasi *email spam*.

1.5. Batasan Masalah

Dalam penelitian ini memiliki titik fokus terhadap pokok masalah. Pokok masalah tersebut yaitu:

1. Dataset yang digunakan berasal dari *UCI Machine Learning* dari penelitian Hewlett-Packard Labs pada tahun 1999.
2. Atribut yang digunakan antara lain, *word_freq_make*, *word_freq_address*, *word_freq_all*, *word_freq_3d*, *word_freq_our*,*word_freq_over*, *word_freq_remove*, *word_freq_internet*, *word_freq_order*, *word_freq_mail*, *word_freq_receive*, *word_freq_will*, *word_freq_people*, *word_freq_report*, *word_freq_addresses*, *word_freq_free*, *word_freq_business* *word_freq_email*, *word_freq_you*, *word_freq_credit*, *word_freq_your*, *word_freq_font*, *word_freq_000*, *word_freq_money*, *word_freq_hp*, *word_freq_hpl*, *word_freq_george*, *word_freq_650*, *word_freq_lab*, *word_freq_labs*, *word_freq_telnet*, *word_freq_857*, *word_freq_data*, *word_freq_415*, *word_freq_85*, *word_freq_technology*, *word_freq_1999*, *word_freq_parts*, *word_freq_pm*, *word_freq_direct*, *word_freq_cs*, *word_freq_meeting*, *word_freq_original*, *word_freq_project*, *word_freq_re*, *word_freq_edu*, *word_freq_table*, *word_freq_conference*, *char_freq_%3B*, *char_freq_%28*, *char_freq_%5B*, *char_freq_%21*, *char_freq_%24*, *char_freq_%23*, *capital_run_length_average*, *capital_run_length_longest* dan *capital_run_length_total*.
3. *Output* yang dihasilkan yaitu *spam* dan bukan *spam*.
4. Skenario uji menggunakan *K-Fold Cross Validation* dengan nilai k= 2, 4, 5, 8 dan 10.

5. Penghitungan jarak terdekat menggunakan *euclidean distance* dengan nilai k tetangga terdekat = 3.