BAB I

PENDAHULUAN

1.1. LATAR BELAKANG MASALAH

Escherichia Coli atau yang sering disingkat *E. Coli* merupakan bakteri gram negatif yang bentuknnya batang pendek (*coccobasil*) dan biasanya atau normalnya hidup berada pada dalam tubuh manusia ataupun hewan berdarah panas yang lebih tepatnya pada saluran pencernaan, bakteri ini bergerak melalui *flagella* [1]. Secara ilmu mikro biologi, bakteri biasanya hidup berkelompok (berkoloni), koloni inilah yang mereka manfaatkan untuk media perkembangbiakan maupun untuk bertahan hidup antar kawanan mereka [2].

Selain itu juga, bakteri *E. Coli* juga dapat ditemui pada jenis daging yang kurang matang dan susu yang tidak dipasteurisasi. Macam-macam bakteri ini dapat menyebar melalui makanan yang tercemar limbah atau kurang higienis dalam proses produksi [3]. Berdasarkan hal berikut, maka metode dari *naïve bayes* inilah yang merupakan metode klasifikasi *data mining* yang mampu memecahkan masalah dari sampel data yang dipakai, hal inilah alasan mengapa metode dari *naïve bayes* menjadi salah satu metode yang ampuh dipakai untuk mengatasi masalah dalam keterbatasan sampel data tersebut [4].

Selain *naïve bayes* dalam penelitian ini juga menggunakan algoritma K-*Nearest Neighbor* (KNN). Dimana prinsip kerja dari K-*Nearest Neighbor* (KNN) yaitu mencari jarak terdekat antara data yang digunakan berdasarkan K tetangga (neighbor) terdekat didalam suatu data pelatihan [5]. Walaupun algoritma KNN adalah algoritma sederhana yang mudah dilakukan, tidak sedikit peneliti yang

meragukan nilai dari k yang dihasilkan. Nilai k yang tinggi dapat mengurangi efek *noise*, tetapi akan membuat hasil prediksi semakin kabur, sedangkan jika nilai k terlalu kecil atau 1, akan mengakibatkan hasil prediksi terasa kaku. Algoritma KNN intinya didasarkan pada klasifikasi terhadap objek berdasarkan data pelatihan yang jaraknya paling dekat dengan objek tersebut [6].

Secara singkat, klasifikasi merupakan pengelompokan objek kedalam kelas tertentu berdasarkan kelompoknya yang biasanya disebut dengan kelas (*class*) [7]. Klasifikasi sendiri merupakan salah satu metode dari *data mining*, yang mana klasifikasi ini merupakan analisis data yang dapat melakukan prediksi sesuai dari label kelas sampel yang akan diklasifikasikan. Klasifikasi ini jugalah yang dapat menghasilkan model-model yang selanjutnya akan digambarkan menurut kelas-kelas yang berada dalam data, metode ini mempunyai beberapa persyaratan yaitu attribut data haruslah yang numerik atau nominal serta label datanya nominal [8].

Klasifikasi melibatkan pengelompokkan bakteri *E. Coli* ke dalam katergori-kategori yang berbeda berdasarkan ciri-ciri yang dimilikinya. Kategori ini merupakan label atau *class* yang berisikan membran sel dan membran plasma pada sel bakteri, yang berperan mengatur zat-zat di dalam dan di luar sel. Maka berdasarkan label inilah dilakukan komparasi dalam prediksi pada *naïve bayes* maupun pengujian jarak terdekat pada kNN untuk identifikasi dan klasifikasi serta pengembangan dalam penelitian bakteri *E. Coli*.

Berdasarkan permasalahan diatas, untuk mendukung upaya tersebut maka dilakukan penelitian yang dituangkan dalam penulisan tugas akhir dengan judul **"Komparasi Penerapan Algoritma *Naïve Bayes* dan KNN Untuk Klasifikasi Bakteri *E. Coli*"**

1.2. RUMUSAN MASALAH

Berdasarkan latar belakang masalah diatas, maka dapat dirumuskan permasalahan dalam penelitian ini yaitu:

1. Bagaimana penerapan (implementasi) algoritma *Naïve Bayes* dan KNN dalam mengklasifikasikan dataset bakteri *E. Coli*?
2. Bagaimana pengujian serta komparasi akurasi antara algoritma *Naïve Bayes* dan KNN apabila digunakan untuk mengklasifikasi bakteri *E. Coli*?

1.3. BATASAN MASALAH

Dalam penelitian ini agar dapat selalu fokus pada pokok permasalahan yang ada, maka dibatasi permasalahan yang akan dibahas sebagai berikut:

1. Dataset yang digunakan untuk klasifikasi adalah dataset bakteri *E. Coli* yang diambil dari situs https://archive.ics.uci.edu/dataset/39/ecoli
2. Pengolahan dataset bakteri *E. Coli* memiliki 8 atribut, yaitu *sequence name* (Nomor aksesi untuk *database* SWISS-PROT), mcg (metode McGeoch), gvh (metode von Heijne), lip (signal von Heijne peptidase II), chg (prediksi banyaknya N-terminus *lipoprotein*), aac (skor analisis asam amino), alm1 (skor/nilai dari membran Alom terprediksi) dan alm2 (skor program Alom tidak terprediksi daerah alm1), serta memiliki 1 label yaitu

class distribution (pembagian kelas). Algoritma yang digunakan yaitu *Naïve Bayes* yang merupakan metode klasifikasi dimana untuk memprediksi probabilitas sebuah *class* dan algoritma KNN untuk mencari jarak terdekat antara data yang akan dievaluasi terhadap K tetangga (*neighbor*) terdekatnya.

3. Tahapan proses yang digunakan merupakan *Knowladge Discovey in Database* (KDD), dimana secara garis besar yaitu *data selection, pre-processing, transformation, data mining* (klasifikasi) dan *interpretation/evaluation*.
4. *Tools* yang digunakan yaitu *RapidMiner* sebagai mesin *data mining* untuk penerapan (implementasi) algoritma *Naïve Bayes* dan kNN serta *Microsoft Excel* sebagai perhitungan evaluasi untuk metode klasifikasi *Naïve Bayes.*

1.4. TUJUAN DAN MANFAAT PENELITIAN

1.4.1. Tujuan Penelitian

Adapun tujuan dari penelitian yang akan dilakukan oleh penulis, yaitu sebagai berikut:

1. Untuk penerapan (implementasi) algoritma *Naïve Bayes* dan KNN untuk klasifikasi dataset bakteri *E.Coli*.
2. Untuk pengujian serta komparasi akurasi antara algoritma *Naïve Bayes* dan KNN untuk klasifikasi dataset bakteri *E.Coli*.

1.4.2. Manfaat Penelitian

Adapun manfaat dari penelitian yang akan dilakukan oleh penulis, yaitu sebagai berikut:

1. Berdasarkan penerapan (implementasi) algoritma *Naïve Bayes* dan KNN untuk klasifikasi dataset bakteri *E.Coli* diharapkan dapat menambah wawasan penulis serta pembaca seputar bakteri *E. Coli* dan keilmuan dalam pengklasifikasian dataset.
2. Hasil pengujian serta komparasi akurasi antara algoritma *Naïve Bayes* dan KNN untuk klasifikasi dataset bakteri *E.Coli* diharapkan dapat mengetahui tingkat akurasi yang baik dari dataset yang diujikan.

1.5 SISTEMATIKA PENULISAN

Untuk mempermudah memahami penulisan laporan tugas akhir ini, maka dibuat sistematika penulisan pada penelitian ini sebagai berikut:

BAB I : PENDAHULUAN

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

BAB II : LANDASAN TEORI

Pada bab ini berisi teori-teori dasar yang mendukung penelitian, dikutip dari buku, jurnal, dan lain-lain yang berfungsi sebagai kerangka atau landasan untuk mendukung pemahaman terhadap penelitian yang peneliti lakukan berupa penjelasan mengenai konsep *data mining* seperti tahapan proses data mining dan

pengelompokkan *data mining*, konsep metode klasifikasi, bakteri *E. Coli*, algoritma *naïve bayes* dan KNN. Pada bab ini juga memuat tinjauan pustaka yang berisi penelitian-penelitian sebelumnya yang berhubungan dengan penelitian ini.

BAB III : METODOLOGI PENELITIAN

Dalam bab metodologi penelitian ini memuat kerangka kerja penelitian, metode pengumpulan data, metode klasifikasi serta alat bantu yang digunakan untuk pembuatan program.

BAB IV : ANALISIS DAN VISUALISASI

Pada bab ini berisi perhitungan analisis menggunakan metode naïve bayes dan KNN terhadap data-data *E. Coli* yang tersedia, hasil dari analisis dan bentuk visualisasi analisis dari tools *RapidMiner* dan *Excel* yang digunakan.

BAB V : PENUTUP

Pada bab penutup terdiri atas kesimpulan dari pembahasan bab-bab sebelumnya serta saran-saran yang berguna bagi perkembangan dengan hasil-hasil penelitian tersebut.