

# BAB I PENDAHULUAN

## 1.1 Latar Belakang

Perkembangan teknologi informasi dan analisis data saat ini berperan penting dalam membantu penelitian biologi dan mikrobiologi, khususnya dalam proses klasifikasi dan identifikasi mikroorganisme. Salah satu bakteri yang sering menjadi objek penelitian adalah *Escherichia coli* (*E. coli*), bakteri gram negatif yang hidup di saluran pencernaan manusia maupun hewan. Meskipun sebagian besar strain *E. coli* bersifat non-patogenik, beberapa di antaranya merupakan penyebab utama penyakit seperti diare, infeksi saluran kemih, dan sepsis. Oleh karena itu, proses identifikasi struktur sel *E. coli* menjadi penting untuk mendukung deteksi dini serta pengembangan terapi berbasis bioteknologi. Penggunaan pendekatan berbasis *machine learning* dalam klasifikasi bakteri memungkinkan peneliti mengelola data biologis dalam skala besar dengan tingkat akurasi yang tinggi [1].

Klasifikasi struktur sel *E. coli* dapat dilakukan menggunakan dataset publik yang disediakan oleh University of California, Irvine (UCI) Machine Learning Repository, yang dikenal sebagai *Ecoli Dataset*. Dataset ini berisi atribut biologis seperti lokasi protein dan karakteristik kimiawi sel yang menjadi dasar analisis. Pendekatan *machine learning* menjadi solusi efektif untuk menangani dataset tersebut karena kemampuannya mengenali pola dan relasi kompleks yang tidak dapat dideteksi melalui analisis konvensional [2]. Dalam konteks ini, algoritma K-Nearest Neighbor (K-NN) sering dipilih karena kesederhanaan implementasi dan efektivitasnya terhadap data non-linier [3].

Metode K-NN bekerja dengan menghitung jarak antar titik data, kemudian menentukan kelas dari suatu data baru berdasarkan mayoritas tetangga terdekatnya. Keunggulan metode ini adalah tidak bergantung pada asumsi distribusi data, sehingga sangat sesuai untuk data biologis dengan variasi tinggi [4]. Dalam penelitian sebelumnya, metode K-NN menunjukkan performa baik dalam

klasifikasi data medis dan biologis, seperti prediksi penyakit ginjal kronik [3], klasifikasi stunting [1], serta prediksi kualitas urin [4]. Penelitian-penelitian tersebut memperkuat dasar teoritis penerapan K-NN pada kasus klasifikasi biologis seperti struktur sel *E. coli*.

Sejumlah penelitian di Indonesia telah menunjukkan efektivitas K-NN untuk berbagai domain klasifikasi. Misalnya, penelitian [5] menggunakan K-NN dalam klasifikasi penyakit tanaman bawang merah dan memperoleh akurasi di atas 90%. Penelitian [6] menerapkan *multi-label K-NN* untuk mengidentifikasi penyakit kanker serviks dengan hasil yang cukup presisi. Sementara itu, penelitian [2] memanfaatkan K-NN untuk klasifikasi ketahanan pangan di Jawa Tengah dan menemukan bahwa penyesuaian nilai  $k$  berpengaruh signifikan terhadap akurasi model. Berdasarkan hasil-hasil tersebut, dapat disimpulkan bahwa K-NN merupakan metode yang fleksibel dan mampu bekerja baik pada berbagai jenis data, termasuk data biologis dan medis.

Meskipun demikian, keberhasilan K-NN sangat bergantung pada pemilihan parameter  $k$  yang tepat dan tahap pra-pemrosesan data yang benar. Nilai  $k$  yang terlalu kecil dapat membuat model terlalu sensitif terhadap noise, sedangkan nilai  $k$  yang terlalu besar dapat menurunkan kemampuan model membedakan pola minoritas [7]. Selain itu, langkah-langkah seperti normalisasi fitur, pembagian data latihan dan uji, serta evaluasi metrik seperti akurasi, presisi, recall, dan F1-score sangat berpengaruh terhadap hasil akhir model [8]. Evaluasi yang komprehensif menggunakan berbagai metrik menjadi penting untuk memastikan performa model tidak bias terhadap kelas dominan.

Penelitian ini bertujuan untuk menerapkan metode K-NN dalam mengklasifikasikan struktur sel *E. coli* berdasarkan dataset dari UCI. Dengan melakukan variasi nilai  $k$  dan mengukur performa menggunakan metrik yang beragam, penelitian ini diharapkan dapat menemukan konfigurasi model terbaik yang seimbang antara bias dan variansi. Selain berkontribusi terhadap literatur ilmiah nasional terkait penerapan *machine learning* dalam bioteknologi, hasil penelitian ini juga diharapkan dapat menjadi acuan pengembangan sistem

klasifikasi otomatis berbasis kecerdasan buatan di laboratorium mikrobiologi. Selain itu, pendekatan ini dapat dikembangkan untuk mendukung sistem deteksi patogen cepat pada sektor medis dan pangan di Indonesia [9].

Selain memberikan kontribusi akademis terhadap penerapan algoritma *machine learning* dalam bidang biologi, penelitian ini juga diharapkan memiliki manfaat praktis dalam mendukung efisiensi proses analisis laboratorium. Dengan adanya model klasifikasi berbasis K-Nearest Neighbor yang mampu mengenali struktur sel *Escherichia coli* secara otomatis, proses identifikasi mikroorganisme dapat dilakukan dengan waktu yang lebih singkat dan tingkat kesalahan yang lebih rendah dibandingkan metode manual konvensional. Penerapan hasil penelitian ini berpotensi membantu pengembangan sistem berbasis kecerdasan buatan yang dapat digunakan dalam bidang kesehatan, industri pangan, maupun penelitian mikrobiologi terapan. Dengan demikian, penelitian ini bukan hanya relevan dalam konteks ilmiah, tetapi juga memberikan nilai tambah nyata bagi kemajuan teknologi dan inovasi di bidang bioteknologi di Indonesia.

## 1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, dapat dirumuskan permasalahan yang akan dikaji dalam penelitian ini adalah :

Bagaimana proses penerapan metode *K-Nearest Neighbor (K-NN)* dalam mengklasifikasikan struktur sel bakteri *Escherichia coli* berdasarkan dataset dari *UCI Machine Learning Repository*, dengan tujuan untuk membangun model klasifikasi yang mampu mengenali pola atribut biologis pada data *E. coli* serta menghasilkan hasil prediksi yang akurat dan konsisten?

## 1.3 Batasan Masalah

Agar penelitian ini berjalan secara fokus dan sesuai dengan tujuan yang telah ditetapkan, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya menggunakan dataset bakteri *Escherichia coli* (*E. coli*) yang bersumber dari *UCI Machine Learning Repository*. Dataset tersebut digunakan untuk proses klasifikasi struktur sel berdasarkan atribut biologis

yang telah disediakan, tanpa melakukan pengambilan data baru atau eksperimen laboratorium tambahan.

2. Metode yang digunakan dalam penelitian ini terbatas pada algoritma *K-Nearest Neighbor (K-NN)* dengan variasi nilai  $k$  tertentu. Penelitian ini berfokus pada penerapan, pengujian, dan analisis pengaruh perubahan nilai  $k$  terhadap hasil klasifikasi, tanpa membandingkan kinerja K-NN dengan metode klasifikasi lain.
3. Evaluasi kinerja model dilakukan dengan menggunakan empat metrik utama, yaitu akurasi, presisi, recall, dan *F1-score*. Analisis hasil difokuskan pada bagaimana metode K-NN menghasilkan tingkat akurasi terbaik serta keseimbangan optimal antara bias dan variansi, tanpa membahas aspek biologis dari struktur sel *E. coli* secara mendalam.

#### **1.4 Tujuan Penelitian**

Berdasarkan batasan masalah yang telah ditetapkan, tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan metode *K-Nearest Neighbor (K-NN)* untuk melakukan klasifikasi terhadap struktur sel bakteri *Escherichia coli* berdasarkan dataset yang diperoleh dari UCI Machine Learning Repository, sehingga dapat membentuk model klasifikasi yang mampu mengenali pola atribut biologis pada data *E. coli*.
2. Mengevaluasi kinerja model K-NN menggunakan metrik akurasi, presisi, recall, dan *F1-score*, untuk menilai sejauh mana metode tersebut mampu memberikan hasil klasifikasi yang akurat dan konsisten.

#### **1.5 Manfaat Penelitian**

Penelitian ini diharapkan dapat memberikan kontribusi yang berarti baik dari sisi teoritis maupun praktis. Adapun manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Manfaat Teoritis

Secara teoritis, penelitian ini diharapkan dapat memberikan kontribusi bagi pengembangan ilmu pengetahuan di bidang *data mining*, *machine learning*, dan bioinformatika, khususnya dalam penerapan algoritma *K-Nearest Neighbor (K-NN)* untuk klasifikasi data biologis yang memiliki karakteristik non-linier. Hasil penelitian ini diharapkan dapat memperkaya literatur terkait pengaruh variasi nilai  $k$  terhadap kinerja model klasifikasi serta memperkuat pemahaman tentang efektivitas metode *supervised learning* sederhana seperti K-NN dalam menganalisis struktur sel mikroorganisme, sehingga dapat dijadikan landasan bagi penelitian lanjutan maupun pengembangan model hibrida di masa mendatang.

## 2. Manfaat Praktis

Secara praktis, penelitian ini diharapkan dapat memberikan manfaat nyata bagi akademisi, praktisi, dan pengembang sistem. Bagi peneliti dan akademisi, hasil penelitian ini dapat dijadikan acuan dalam mengembangkan sistem klasifikasi berbasis *machine learning* di bidang bioteknologi dan mikrobiologi. Bagi praktisi laboratorium, penelitian ini dapat membantu mempercepat proses identifikasi struktur sel *Escherichia coli* secara komputasional dengan tingkat akurasi tinggi, sehingga meningkatkan efisiensi analisis data biologis. Selain itu, bagi pengembang sistem kecerdasan buatan, hasil penelitian ini dapat menjadi dasar untuk merancang aplikasi otomatis yang mampu mengklasifikasikan data biologis dengan lebih cepat dan andal.

## 1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memudahkan pembaca dalam memahami alur penelitian, dan terbagi menjadi beberapa bab sebagai berikut:

### BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

### BAB II TINJAUAN PUSTAKA

Bab ini membahas teori-teori yang mendukung penelitian, termasuk konsep dasar K-Nearest Neighbor (K-NN), klasifikasi struktur sel *Escherichia coli*, dan penelitian terdahulu yang relevan.

### BAB III METODE PENELITIAN

Bab ini menjelaskan metode penelitian yang digunakan, mulai dari jenis dan sumber data, langkah-langkah pra-pemrosesan, penerapan metode K-NN, pemilihan parameter, serta teknik evaluasi performa klasifikasi.

### BAB IV HASIL DAN PEMBAHASAN

Menyajikan hasil penelitian, termasuk analisis performa K-NN, pengaruh parameter  $k$  dan fungsi jarak

### BAB V PENUTUP

Berisi kesimpulan dari hasil penelitian yang telah dilakukan dan saran untuk pengembangan penelitian selanjutnya.