BAB 1
PENDAHULUAN

1.1 Latar Belakang

Bioinformatika merupakan ilmu dari biologi, matematika, dan ilmu komputer untuk menganalisis data-data biologi. Bioinformatika mempelajari data biologi dengan menerapkan teknik komputasional untuk pengelolaan dan penganalisissian informasi biologis. Data-data tersebut disimpan secara digital ke NCBI (*National Center for Biotechnology Information*) [1]. Bioinformatika menggunakan teknologi komputer untuk mengidentifikasi dasar genetik penyakit, adaptasi unik suatu sel, sifat sel yang diinginkan, atau perbedaan antar populasi. Beberapa pengaplikasian dari bioinformatika dapat membantu bidang pengobatan seperti mengidentifikasi korelasi antara urutan gen dan penyakit, untuk memprediksi struktur protein dari urutan asam amino yang dapat membantu dalam desain obat baru dan menyesuaikan perawatan untuk setiap pasien yang berbeda [2].

Dengan menerjemahkan suatu molekul menjadi sebuah berkas yang dapat diolah komputer, untuk memastikan data bioinformatika tersedia untuk masyarakat umum, data bioinformatika diunggah ke dalam *database* yang dapat diakases via internet. Beberapa *website* popular yang menyediakan *database* bagi pengirim data bioinformatika adalah Genbank dari NCBI (*National Center for Biotechnology Information*), EMBL (*European Molecular Biology Laboratory*), SwissProt (*Swiss Institute of Bioinformatics),* PIR (*Protein Information Resource*) dan DDBJ (*DNA Database of Japan*). Repositori penyedia *database* tersebut memperbaharui *database* nya setiap hari. Sedangkan, SwissProt dan PIR yang menyediakan data tersebut dibuat dan diteliti oleh para ahli. Salah satu pertanyaan kritis yang muncul dalam penggunaan *database* adalah keandalan informasinya. Bagi *database* yang diteliti oleh para ahli, ada alasan kuat untuk percaya bahwa sumber informasi harus dapat dipercaya. Karena hanya pengirim yang dapat mengubah entri, hal yang salah dan atau membingungkan dapat tersimpan di *database* selama bertahun-tahun [3]. Walaupun pengumpulan data adalah hal yang diperlukan, informasi yang disimpan dalam *database* ini pada dasarnya tidak berguna sampai dianalisis.

Keakuratan data dibutuhkan setinggi-tingginya untuk analisis data. Oleh karena itu dalam penelitian ini akan menggunakan metode *data mining* dengan model klasifikasi *Naïve Bayes* dan J48 untuk kumpulan data protein SARS-CoV-2 yang disediakan oleh *database website* NCBI dan menganalisis hasil akurasi antara model klasifikasi *Naïve Bayes* dan J48.

1.2 Rumusan Masalah

Berdasarkan latar belakang pada bagian 1.1, dapat diuraikan rumusan masalahnya, berupa :

1. Bagaimana mencari akurasi pada kumpulan data protein virus SARS-CoV-2?
2. Bagaimana memprediksi protein virus SARS-CoV-2?

1.3 Tujuan Pembahasan

Berdasarkan rumusan masalah pada bagian 1.2, berikut diuraikan dan dijabarkan garis besar hasil pokok yang ingin dicapai setelah permasalahan dijawab untuk menjelaskan :

1. Mencari model klasifikasi dengan akurasi yang tepat untuk kumpulan data.
2. Memprediksi protein virus SARS-CoV-2 dengan metode *data mining* J48 dan *Naïve Bayes*.

1.4 Ruang Lingkup

Adapun ruang lingkup permasalahan yang akan dibahas dalam penulisan laporan ini berupa penjelasan:

1. Penjelasan Bioinformatika.
2. Pengambilan kumpulan data dari *website* NCBI.
3. Penerapan *data mining* algoritma *Naïve Bayes* dan J48 dengan aplikasi WEKA.
4. Menganalisis hasil detail akurasi dan berapa lama waktu yang dibutuhkan dalam pembuatan model dan prediksi data antara algoritma *Naïve Bayes* dan J48.
5. Pengumpulan data latih diunduh pada tanggal 13 Maret 2021 dan data uji pada tanggal 13 April 2021.
6. Perangkat Keras yang digunakan Berikut adalah spesifikasi perangkat keras yang digunakan dalam penelitian ini:
    - *Operating System: Windows 10 Home 64-bit (10.0, Build 19041)*
    - *System Manufacturer: ASUSTeK COMPUTER INC.*

- *System Model: GL553VD*
- *Processor: Intel® Core™ i7-7700HQ CPU @ 2.80GHz (8 CPUs), ~2.8GHz*
- *Memory: 16384MB RAM*

1.5 Sumber Data

Sumber data dalam laporan Analisis Akurasi Data Protein Virus SARS-CoV-2 dengan menggunakan Metode *Data mining* ini menggunakan sumber data primer dan data sekunder. Data primer berupa kumpulan data protein virus SARS-CoV-2 yang sudah dibersihkan dalam tahap praproses data, sedangkan data sekunder yang diambil adalah jurnal, artikel, dan buku yang membantu pembuatan laporan ini.

1.6 Sistematika Penyajian

BAB I Pendahuluan

Bab ini menjelaskan pendahuluan pada laporan tugas akhir yang berupa latar belakang, rumusan masalah, tujuan pembahasan, ruang lingkup dan sistematika penyajian.

BAB II Kajian Teori

Bab ini menjelaskan teori yang mendukung laporan tugas akhir ini dan penjelasan mengenai teori tersebut dan sejarahnya.

BAB III Metodologi Penelitian

Bab ini menjelaskan tahapan penelitian pada laporan ini yang diilustrasikan sebagai bagan juga penjelasannya.

BAB IV Hasil dan Pembahasan

Bab ini menjelaskan hasil penelitian mulai dari proses penelitian pengumpulan data sampai analisis hasil data.

BAB V Penutup

Bab ini menjelaskan kesimpulan penelitian dan saran untuk peneliti berikutnya.