BAB I

PENDAHULUAN

1.1. Latar Belakang

Data ekspresi gen atau DNA *microarray* merupakan data berdimensi tinggi yang menjelaskan proses mengenai penguraian DNA sehingga menghasilkan protein. Ekspresi gen sangat banyak manfaatnya salah satunya adalah untuk mengetahui jenis perkembangan penyakit dalam tubuh manusia. Salah satu penyebab atau masalah yang signifikan pada data berdimensi tinggi ini seperti data yang tidak relevan, terjadinya eror pada data (*outlier*) tidak kompleksitas dalam membangun klasifikasi dan beberapa data yang *missing value*[8]. Sehingga permasalahan yang terjadi pada data ekspresi gen memerlukan biaya dan waktu yang sangat tinggi dalam perhitungan komputasi. Semakin banyak masalah yang terjadi maka semakin sulit untuk diselesaikan. Untuk menangani segala permasalahan yang ada pada data ekspresi gen diperlukan suatu metode komputasi yaitu *data mining*. *Data mining* merupakan proses penggabungan dari berbagai bidang ilmu seperti teknik matematika, statistika, kecerdasan buatan dan *Machine learning* yang berfungsi untuk mengelola suatu informasi pada data dalam jumlah yang sangat besar.

Penelitian data ekspresi gen sebelumnya sudah pernah dilakukan pada salah satu jurnal yang telah dipublikasi yaitu “*Review on Feature Selection of Gene Expression Data for Autism Classification*”[8]. Penelitian tersebut menjelaskan mengenai data ekspresi gen dengan menggunakan *Genetic Algorithm*. Pada tugas akhir ini permasalahan yang diselesaikan adalah mengenai data ekspresi gen dengan menggunakan algoritma *K-means Clustering* yang digunakan sebagai klasifikasi pada data, dan untuk menghasilkan suatu nilai akurasi yang lebih baik. Salah satu kelebihan menggunakan *k-means clustering* ini yaitu mudah untuk diimplementasikan dan proses perhitungannya relative cepat. Selain algoritma *k-means clustering* penulis juga menggunakan salah satu algoritma yang dikembangkan oleh *PCA* yaitu *Weighted Principal Component Analysis* yang digunakan sebagai fitur seleksi pada data berdimensi tinggi.

1.2. Perumusan Masalah

Berdasarkan dari permasalahan untuk mengolah data ekspresi gen pada penyakit kanker ini adapun rumusan masalah sangat spesifik yaitu :

- a. Bagaimana meimplementasikan dari fitur seleksi *Weighted PCA* dan algoritma *K-Means Clustering* dalam mengklasifikasikan data ekspresi gen penyakit kanker
- b. Bagaimana hasil performansi yang didapatkan oleh *K-means* dan *WPCA* dalam mengklasifikasikan data ekspresi gen penyakit kanker.

Batasan Masalah

Adapun Batasan masalah yang diangkat pada tugas akhir ini adalah :

- a. Data yang digunakan adalah data ekspresi gen penyakit *Leukimia*
- b. Data tersebut di peroleh dari *Kent Ridge Bio-medical Data Set Repository*.

1.3. Tujuan

Berdasarkan dari permasalahan untuk mengolah data ekspresi gen pada penyakit kanker ini adapun tujuan yang sangat spesifik yaitu :

- a. Meimplementasikan dari fitur seleksi *Weighted PCA* dan algoritma *K-Means Clustering* dalam mengklasifikasikan data ekspresi gen penyakit kanker
- b. Mengukur hasil performansi yang didapatkan oleh *K-means* dan *WPCA* dalam mengklasifikasikan data ekspresi gen penyakit kanker.

1.4. Sistematika Penulisan

Sistematika penulisan pada tugas akhir ini adalah :

1. BAB 1 PENDAHULUAN

Pada bagian ini akan dijelaskan permasalahan serta latar belakang dari tugas akhir ini

2. BAB 2 KAJIAN PUSTAKA

Pada bagian ini akan dijelaskan metode serta teori yang digunakan untuk menyelesaikan permasalahan pada penyakit data ekspresi gen

3. BAB 3 PERANCANGAN SISTEM

Pada bagian ini akan dijelaskan kinerja sistem dan gambaran umum dari proses pengolahan data ekspresi gen

4. **BAB 4 PENGUJIAN DAN ANALISIS**

Pada bagian ini akan dijelaskan pengujian sistem serta analisis dari hasil pengujian

5. **BAB 5 PENUTUP**

Pada bagian ini akan diambil sebuah kesimpulan dari keseluruhan permasalahan yang telah diselesaikan.