1. PENDAHULUAN

1.1 Latar Belakang

Data mining memiliki banyak sekali manfaat dalam berbagai masalah pengolahan data, sehingga data-data yang tidak memiliki informasi penting dapat digali dan dianalisa. Dalam prosesnya data yang memiliki banyak dimensi atau bisa disebut juga sebagai data dimensi tinggi dapat mengakibatkan proses penggalian informasi menjadi lebih sulit. Data dimensi tinggi memiliki dimensi yang sangat banyak yang jumlah dimensinya mencapai ratusan bahkan ribuan dimensi, sehingga kompleksitas dari data tersebut menjadi sangat besar. Tidak menutup kemungkinan bahwa dari keseluruhan dimensi pada data, sebenarnya ada dimensi yang tidak perlu digunakan pada proses *data mining*.

Fenomena yang berkaitan dengan permasalahan data dimensi tinggi ini biasanya disebut juga sebagai fenomena *Curse of Dimensionality*. *Curse of Dimensionality* atau kutukan dimensi dapat membuat proses pengolahan data menjadi kurang efektif dan efisien, sehingga diperlukan teknik tertentu untuk mereduksi dimensi sehingga memiliki tingkat akurasi (performansi) yang baik [4]. Selain itu, mereduksi dimensi juga bertujuan untuk memudahkan pengolahan data yang dilakukan data mining dan membuat model klasifikasi lebih mudah untuk dipahami[12].

Penggabungan antara algoritma *data mining* dengan algoritma *Evolutionary Algorithms* (EAs) menjadi salah satu solusi untuk mengatasi permasalahan yang berkaitan dengan fenomena kutukan dimensi. EAs adalah algoritma-algoritma optimasi yang yang berbasis evolusi biologi yang ada pada dunia nyata. EAs bekerja dengan cara membangkitkan, menguji, dan berusaha memperbaiki sekumpulan kandidat solusi sampai ditemukan satu solusi yang bisa diterima [2]. Algoritma EAs yang digunakan pada tugas akhir ini adalah *Genetic Algorithm* (GA).

K-Nearest Neighbour merupakan merupakan salah satu metode *supervised learning* yang ada pada data mining. Algoritma ini biasanya digunakan sebagai teknik klasifikasi dan regresi[5]. Ketepatan klasifikasi pada algoritma KNN sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi.

Algoritma *K-Nearest Neighbour* (KNN) merupakan algoritma data mining yang dapat digunakan untuk melakukan klasifikasi dan *Genetic Algorithm* (GA) yang dapat membantu dalam memaksimalkan akurasi klasifikasi dari atribut sehingga tingkat keakuratan (performansi) menjadi lebih optimal[10]. Penelitian dengan menggunakan Algoritma KNN dengan GA untuk data berdimensi tinggi sudah pernah dilakukan sebelumnya pada jurnal *Classification of Heart Disease Using K-Nearest Neighbour and Genetic Algorithm* [1].

1.2 Rumusan Masalah

Rumusan masalah dari tugas akhir ini adalah sebagai berikut :

1. Bagaimana mengimplementasikan algoritma KNN dan algoritma genetika untuk memprediksi data penyakit berdimensi tinggi?
2. Bagaimana performansi dari algoritma KNN dan GA dalam memprediksi penyakit dengan data berdimensi tinggi?

Adapun batasan masalah dari tugas akhir ini adalah sebagai berikut :

1. Data yang digunakan merupakan data beberapa penyakit, antara lain penyakit *Leukimia* dan *Colon Tumor* yang diambil dari *Biomedical Dataset* pada Kent Ridge [6].
2. Tidak ada penanganan *outlier* dan *missing value* pada data.
3. Perhitungan jarak data uji ke data latih pada algoritma KNN menggunakan metode *Euclidean distance*.
4. KKN digunakan sebagai evaluasi individu pada GA.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah sebagai berikut :

1. Mengimplementasikan algoritma KNN dan algoritma genetika untuk memprediksi data penyakit berdimensi tinggi.
2. Mengetahui performansi dari algoritma KNN dan GA dalam memprediksi penyakit dengan data berdimensi tinggi.

1.4 Sistematika Penulisan

Sistematika penulisan yang digunakan pada penulisan tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini berisikan latar belakang masalah, perumusan masalah, tujuan, hipotesis, dan sistematika penulisan. Secara umum bab ini memberi gambaran umum tentang sistem yang akan dibangun.

BAB II KAJIAN PUSTAKA

Bab ini memberikan penjelasan teori yang terkait dengan metode-metode yang akan digunakan dalam analisis penyelesaian masalah. Penjelasan teori yang akan digunakan berkaitan erat dengan data mining, data dimensi tinggi, algoritma KNN serta GA.

BAB III METODOLOGI DAN DESAIN SISTEM

Bab ini memberi gambaran serta tahapan-tahapan dari sistem yang akan dibangun. Pada bab ini berisikan rancangan sistem yang akan dibangun, dataset yang digunakan, langkah *preprocessing* data serta skenario pengujian sistem.

BAB IV PENGUJIAN DAN ANALISIS

Bab ini berisikan pembahasan hasil pengujian dan analisis dari sistem yang dibangun berdasarkan skenario pengujian yang dibuat.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil pengujian sistem yang telah dibahas pada bab sebelumnya. Serta saran-saran yang diperlukan untuk pengembangan lebih lanjut.