1. Pendahuluan

1.1. Latar Belakang

Dengan begitu banyaknya tantangan yang dimiliki oleh *data mining*, pengolahan data dalam berbagai macam penelitian menjadi sangat sulit untuk ditangani, tantangan tersebut meliputi [3]: *dimensionality, complex and heterogeneous data, data quality, data ownership and distribution, privacy preservation,* dan *streaming data.* Tantangan tersebut kemudian menimbulkan salah satu masalah dalam industri kesehatan [11] yaitu data penyakit berdimensi tinggi, kategori tantangan ini adalah *dimensionality.*

Dalam kasus ini, data tersebut memiliki jumlah atribut yang begitu banyak, semakin banyak atribut maka semakin banyak memakan waktu dan memanfaatkan upaya komputasi yang berlebihan sehingga data sulit untuk ditangani [5]. Maka dari itu, untuk menjawab tantangan dari masalah tersebut reduksi dimensi sangat diperlukan agar mendapatkan nilai akurasi yang lebih baik untuk setiap penelitian.

Reduksi dimensi ini dilakukan untuk mengurangi bagian-bagian atribut yang tidak diperlukan pada data yang sedang diteliti, contohnya: terdapat nilai dari salah satu atribut tidak ada dalam data (*missing value)* hal ini mungkin sekali terjadi dikarenakan informasi yang diperlukan tidak dapat diperoleh. Selain itu, terdapat juga atribut yang karakteristiknya berbeda diantara atribut-atribut lain (*outlier*) dan adanya *error* pada data (*noise*). Itulah gangguan-gangguan yang terdapat pada data dikarenakan kesalahan dalam pengumpulan informasi.

Data mining menggunakan *Evolutionary Algorithms* (EAs) atau bisa disebut sebagai *Evolutionary data mining* dapat digunakan untuk permasalahan tersebut. EAs dapat membantu *data mining* untuk dapat mereduksi dimensi dengan cara membuat aturan-aturan secara acak yang kemudian akan diseleksi untuk mendapatkan atribut-atribut yang paling optimal. Banyak studi yang membahas tentang *evolutionary data mining* diantaranya: penggunaan *evolutionary data mining* pada penyakit diabetes [11], kumpulan data medis [5], prediksi penyakit jantung [9] dan masih banyak lagi.

Algoritma *K-Means* sebagai *instance reduction* dapat digunakan untuk menghilangkan gangguan terhadap data dan *Genetic Algorithm* (GA) digunakan untuk memilih atribut yang paling optimal dalam data tersebut. *Support Vector Machine* (SVM) digunakan sebagai *tools* untuk melakukan klasifikasi berdasarkan keluaran yang telah diperoleh oleh GA untuk mendapatkan akurasi yang lebih baik[5,11]. Sebelumnya penelitian terhadap *K-Means,* GA, dan SVM ini pernah dilakukan dalam memprediksi penyakit diabetes dan memperoleh akurasi sebesar 98% [11]. Pada tugas akhir ini dilakukan prediksi penyakit menggunakan algoritma *K-Means,* GA dan SVM pada kasus data berdimensi tinggi.

1.2. Perumusan Masalah

Permasalahan yang dapat diselesaikan dalam tugas akhir ini terdiri dari:

1. Bagaimana implementasi *K-Means* dan GA dengan mengintegrasikan SVM pada data penyakit berdimensi tinggi ?
2. Bagaimana cara kerja *K-Means* untuk dapat menghilangkan gangguan data dan GA untuk memilih atribut paling yang optimal pada data penyakit berdimensi tinggi ?
3. Bagaimana prediksi yang didapatkan SVM pada data penyakit berdimensi tinggi ?
4. Bagaimana performansi yang didapatkan SVM pada data penyakit berdimensi tinggi ?

1.3. Batasan Masalah

Adapun batasan masalah dari tugas akhir ini, yaitu :

1. *Data set* yang digunakan untuk tugas akhir ini adalah data penyakit berupa data ekspresi gen yang berasal dari *Kent Ridge Bio-medical Data Set Repository* [6].
2. Algoritma *K-Means Clustering* digunakan sebagai algoritma untuk mereduksi *record*.
3. Algoritma EAs yang digunakan adalah GA yang berfungsi sebagai *feature selection.*
4. SVM digunakan sebagai klasifikasi.

1.4. Tujuan

Tujuan untuk menyelesaikan masalah tersebut adalah :

1. Mengimplementasikan Algoritma *K-Means* dan GA untuk mereduksi dimensi dengan mengitegrasikan SVM pada data penyakit berdimensi tinggi.
2. Mengetahui cara kerja *K-Means* dan GA pada data penyakit berdimensi tinggi.
3. Mengetahui hasil prediksi SVM pada data penyakit berdimensi tinggi.
4. Menganalisis hasil performansi yang didapatkan SVM pada data penyakit berdimensi tinggi.

1.5. Sistematika Penulisan

Pada sub bab ini dijelaskan secara singkat mengenai uraian lingkup isi dari setiap bab yang ada dalam buku tugas akhir ini. Bab 1 menjelaskan tentang latar belakang masalah berdasarkan judul yang diajukan penulis serta tujuan dari pembuatan tugas akhir ini. Bab 2 menjelaskan tentang teori-teori pendukung berdasarkan masalah yang dihadapi. Bab 3 menjelaskan tentang perancangan sistem dan skenario yang dibuat untuk penyelesaian masalah yang dihadapi. Bab 4 menjelaskan tentang hasil pengujian dan analisis sistem yang telah dibuat. Bab 5 berisi kesimpulan dan saran atas selesainya tugas akhir ini.