1. Pendahuluan

1.1 Latar belakang

Perkembangan data yang pesat tidak lepas dari perkembangan tekhnologi informasi yang memungkinkan data dalam jumlah besar terakumulasi, ledakan data hampir terjadi di setiap penjuru dunia baik industri, instansi dan internet. Dengan kondisi seperti ini terdapat banyak tuntutan untuk menemukan informasi berguna yang tenggelam dalam tumpukan data dari berbagai sumber. Data dengan jumlah yang begitu besar ini akan sangat menyulitkan apabila kita ingin menganalisa apakah terdapat suatu kesalahan dalam data tersebut. Data yang mempunyai sifat dan karakteristik yang berbeda dari data pada umumnya dan mempunyai kemunculan kejadian relatif sedikit dikatakan sebagai *outlier*.

Outlier detection dapat didefinisikan sebagai pencarian terhadap sebagian kecil dari data, yang memiliki sifat yang berbeda jika dibandingkan dengan data keseluruhan. *Outlier* sendiri dapat didefinisikan sebagai sebuah titik data pada suatu basis data dimana sangat berbeda dibandingkan dengan titik data pada basis data pada umumnya. *Outlier* seringkali mempunyai informasi yang sangat berguna karena memiliki karakteristik yang tidak normal, sehingga dapat menghidupkan kecurigaan misalnya pada aplikasi kecurangan kartu kredit, *network intrusion detection*, aplikasi keuangan dan lain lain.

Dalam data mining terdapat banyak metoda dalam pencarian *outlier* seperti *clustering* yang mendefinisikan sebuah *outlier* tidak terdapat dalam *cluster* tersebut, dengan kata lain, *clustering* secara implisit mendefinisikan *outlier* sebagai *noise* dari suatu cluster tertentu. Teknik lainnya metode statistik dengan medefinisikan sebuah *outlier* berada diluar sekumpulan data yang ada. Metode *distance-based* mendefinisikan sebuah *outlier* berada jauh dari pusat data. Metode *density-based* mendefinisikan sebuah *outlier* merupakan sekumpulan titik data dengan kepadatan yang sangat rendah[15].

Permasalahannya adalah metoda-metoda yang ada seperti metode *clustering*, *m*etode *distance-based* dan metode *density-based* tidak mengkhususkan pencarian outlier pada data yang yang bersifat *categorical*, malah metoda tersebut kebanyakan untuk menangani data yang bersifat *numeric*, padahal dalam aplikasi didunia nyata terdapat data yang bersifat *categorical* bukan hanya *numeric*. Maka dengan metoda *LSA* algoritma akan memecahkan problem yang ada pada aplikasi di dunia nyata yang banyak mengandung data yang bersifat *categorical*[8].

1.2 Perumusan masalah

Masalah pokok yang akan diteliti adalah :

1. Bagaimana cara menguji algoritma *LSA* pada pendeteksian *outlier* pada *categorical* data.
2. Bagaimana analisa nilai perhitungan nilai *entropy* untuk mendeteksi *outlier*.
3. Bagaimana implementasi algoritma *LSA* pada aplikasi *outlier detection*.
4. Bagaimana analisa akurasi algoritma *LSA* dalam menentukan *outlier*.

Dalam penyusunan tugas akhir ini permasalahan dibatasi dalam beberapa hal, yaitu :

1. Algoritma yang digunakan adalah *Local Search Algorithm* (LSA).
2. Tidak menangani *preprosesing* data.
3. Tipe data berupa *categorical.*
4. Dataset bersifat *supervised.*
5. Input data yang akan dilakukan deteksi *outlier* adalah file *.arff.*
6. Pengujian dilakukan pada dataset dengan distribusi kelas yang *imbalanced* dan telah diketahui jumlah outliernya.
7. Analisis performansi meliputi akurasi obyek *outlier* yang dihasilkan, dan waktu eksekusi terhadap peningkatan jumlah data dan penambahan jumlah inputan k-outlier.

1.3 Tujuan

Tujuan dari penelitian tugas akhir ini adalah sebagai berikut :

1. Implementasi Algoritma *LSA* untuk mendeteksi *outlier* pada data *categorical*.
2. Membangun perangkat lunak deteksi *outlier* dengan menerapkan metode *LSA*.
3. Melakukan analisa terhadap perangkat lunak untuk menguji akurasi dan waktu deteksi terhadap penambahan jumlah data dan nilai inputan *k-outlier*.

1.4 Metodologi penyelesaian masalah

Metodologi yang akan digunakan dalam merealisasikan tujuan dan pemecahan masalah di atas adalah dengan menggunakan langkah-langkah berikut.

1. Studi pustaka
   Pada tahap ini dipelajari metoda LSA dalam pendektesian *outlier* pada *categorical* data, dan metoda – metoda lainnya seperti metode *clustering, m*etode *distance-based* dan metode *density-based*.
2. Analisis dan Desain
   Pada tahap ini dilakukan analisis pemecahan dari permasalahan pada data *categorical* yang akan dideteksi *outlier*-nya dengan menggunakan dasar teori yang telah dipelajari pada tahap sebelumnya.
3. Implementasi
   Hasil yang telah dilakukan pada tahap perancangan dapat diimplementasikan pada bahasa pemrograman.
4. Pengujian dan Evaluasi
   Menganalisis dengan mengukur akurasi pendeteksian *outlier* hasil implementasi metoda *LSA*.