1. Pendahuluan

1.1. Latar Belakang

Clustering adalah salah satu teknik yang terdapat dalam data mining. *Clustering* adalah suatu proses pengelompokkan obyek baik fisik maupun abstrak ke dalam suatu kelas atau *cluster* yang berisi kumpulan obyek yang *similarity* nya tinggi. [2]

Permasalahan pada metode *clustering* yang telah ada adalah kebanyakan hanya dapat diterapkan pada data yang bersifat numerik, karena untuk *clustering* pada data numerik relatif lebih mudah dalam penetapan *similarity* dari *geometric position* nya [1]. Sedangkan untuk data yang bersifat kategorikal, yaitu data yang mempunyai nilai suatu himpunan kategori [3], misal pada atribut *color*, *value* nya adalah *blue*, *white*, *black*, akan sulit diukur nilai *similarity* nya, sehingga dapat terjadi kesalahan dalam memasukkan data tersebut dalam suatu *cluster*. Maka dari itu, pada data kategorikal, akan digunakan metode *clustering* yang berbeda dengan data yang bersifat numerik.

Terdapat beberapa algoritma *clustering* untuk data kategorikal yang telah dirancang sebelumnya seperti CACTUS dan ROCK, namun pada penerapannya di dunia nyata, algoritma-algoritma tersebut masih belum dapat menangani *uncertainty* (ketidakpastian) pada proses clustering. Pada perkembangannya, telah di desain metode *clustering* yang menangani *uncertainty* ini dengan menerapkan himpunan *fuzzy* pada proses *clustering* nya seperti pada fuzzy k-modes dan fuzzy centroids, namun algoritma tersebut membutuhkan beberapa kali pengujian agar menghasilkan nilai parameter yang tepat sehingga dapat mencapai kondisi stabilitas yang ingin dicapai.

Terdapat suatu algoritma yang dirancang untuk dapat memecahkan masalah diatas, yaitu Min-Min Roughness (MMR). Algoritma MMR ini dirancang agar dapat menangani *uncertainty* pada saat proses *clustering* data kategorikal dengan menerapkan penggunaan Rough Set Theory (RST). RST adalah suatu metode untuk mendapatkan *decision making* pada kemunculan *uncertainty* dengan memakai konsep *lower approximation* dan *upper approximation* yang selanjutnya digunakan untuk mendapatkan nilai *roughness* pada tiap atribut terhadap atribut-atribut lainnya. Selanjutnya, algoritma MMR mencari nilai *roughness* yang minimal pada tiap-tiap atribut tersebut, dan memproses ke seluruh atribut yang ada sehingga didapatkan nilai *roughness* minimal untuk tiap-tiap atribut berdasarkan seluruh atribut lainnya. Langkah selanjutnya adalah melakukan pembagian (*split*) terhadap kumpulan objek sampai didapatkan jumlah *cluster* yang sesuai dengan parameter.

Pada tugas akhir ini akan mengimplementasikan algoritma Min-Min Roughness (MMR) untuk melakukan *clustering* pada data kategorikal yang berbasis Rough Set Theory (RST) dan menganalisis nilai akurasi dari hasil yang didapatkan.

1.2. Perumusan Masalah

Beberapa masalah yang dapat diangkat dari tugas akhir ini adalah :

1. Bagaimana penerapan algoritma MMR ini dalam menyelesaikan permasalahan *clustering* data kategorikal ?
2. Bagaimana paramater jumlah cluster mempengaruhi nilai *purity* dari hasil *clustering* ?

Batasan masalah dari tugas akhir ini adalah :

1. *Dataset* yang digunakan berasal dari UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html.
2. *Dataset* bertipe kategorikal dan mempunyai informasi di kelas mana tiap objek berada.

1.3. Tujuan

Tujuan dari tugas akhir ini adalah

1. Mengimplementasikan algoritma MMR dalam proses *clustering* data kategorikal.
2. Mengevaluasi nilai *purity* dari hasil proses *clustering* berdasarkan jumlah cluster yang terbentuk.

1.4. Metodologi Penyelesaian Masalah

Metodologi yang digunakan pada tugas akhir ini adalah :

1. Studi Literatur
   Pencarian sumber-sumber dan referensi literatur yang berkaitan dengan data mining, *clustering*, data kategorikal, algoritma MMR dan Rough Set Theory.
2. Pengumpulan Data
   Mempersiapkan *dataset* berupa data bertipe kategorikal yang akan digunakan dalam proses *clustering.*
3. Perancangan dan Implementasi Sistem
   Perancangan sistem yang akan diimplementasikan yaitu penggunaan algoritma MMR dalam proses *clustering* pada data kategorikal berbasis Rough Set Theory, penentuan bahasa pemrograman yang akan digunakan, beserta fungsionalitas serta antarmuka.
4. Pengujian dan Analisis Hasil
   Pengujian dari sistem yang telah diimplementasikan dengan data yang ada selanjutnya dilakukan analisis terhadap jumlah *cluster* yang terbentuk dan nilai *purity* dari hasil proses *clustering* tersebut.
5. Penyusunan Laporan Tugas Akhir
   Penyusunan laporan dan dokumentasi dalam pembuatan tugas akhir sesuai dengan kaidah dan tata cara yang telah ditetapkan.