

# 1. Pendahuluan

## 1.1 Latar belakang

Perkembangan teknologi dalam hal penyimpanan dan pengolahan data pada saat ini sangat berkembang pesat, hal ini dikarenakan oleh kebutuhan manusia akan informasi yang cepat dan tepat. Oleh itu, pada saat ini sebagian besar organisasi atau perusahaan menyimpan informasi transaksi yang berlangsung di dalam organisasi tersebut, hal ini juga didukung oleh murahnya penerapan teknologi. Data yang banyak yang tersimpan di dalam *storage* perusahaan tidaklah berarti apa-apa jika tidak diolah dan digunakan. Kebutuhan akan analisis data tersebut mendorong penerapan dari berbagai teknik analisis data dari berbagai bidang ilmu seperti statistika, kecerdasan buatan, database dan lain-lain untuk menganalisis data tersebut agar dapat digunakan. Penggabungan analisis dari berbagai bidang ilmu tersebut dikenal dengan *data mining*.

*Data mining* bertujuan untuk menemukan pola yang ada dalam suatu kumpulan data. Pola yang didapat dalam *data mining* dapat digunakan sebagai pertimbangan dalam pengambilan keputusan dalam instansi atau perusahaan sumber data tersebut. Salah satu teknik dalam *data mining* adalah klasifikasi. Klasifikasi merupakan proses menemukan model atau fungsi untuk dapat memprediksi suatu objek yang belum diketahui kelasnya.

Dalam klasifikasi, data yang digunakan untuk proses pembentukan pola sangatlah penting. Untuk mendapatkan pola yang lebih baik dibutuhkan data yang baik pula. Pada klasifikasi ada dua proses utama yaitu proses *training* dan proses *testing*. Proses klasifikasi digunakan variabel input untuk *training* dan *testing* untuk mengetahui tingkat akurasi kebenaran model yang dibangun. Pemilihan variabel yang digunakan untuk proses klasifikasi sangatlah penting. Tujuan pemilihan variabel adalah untuk meningkatkan performansi prediksi suatu kelas, harga prediksi yang efektif, kemudahan visualisasi, pemahanan data dan juga untuk mengurangi dimensionalitas dari variabel input.

Pada tugas akhir ini diimplementasikan pemilihan variabel sebelum proses klasifikasi dengan menggunakan *Rough Set Approach*. *Rough Set Approach* merupakan teknik pendefinisian kelas yang ekivalen dalam *dataset* secara kasar (*roughly*). *Rough Set Approach* bertujuan untuk menemukan minimal subset dari variabel yang akan digunakan pada proses *learning*. Himpunan minimal variabel dari suatu *dataset* dibuat dengan memilih minimal subset dari *dataset* asal. Himpunan variabel yang dapat mewakili beberapa *rule* dinamakan *core*. Setelah seluruh *core* yang dapat mewakili *rule* yang lain didapatkan, maka *dataset* yang baru didapatkan dengan menghapus variabel yang tidak termasuk dalam *core*, setelah akan didapat *dataset* baru yang sebagai hasil dari *variable selection*.

## 1.2 Perumusan masalah

Klasifikasi dalam *data mining* digunakan untuk mencari pola dengan cara menganalisis sekumpulan *dataset* yang mendeskripsikan dan membedakan kelas-kelas data. Tujuan dari pembentukan pola adalah untuk memprediksi data yang

belum diketahui kelasnya. Pada klasifikasi, *dataset* akan dijadikan sebagai data input untuk proses *training* dan *testing*. Namun untuk jumlah data yang besar dan variabel yang banyak, sangat kompleks untuk membuat model *classifier* dari *dataset* tersebut. Oleh karena itu, sebelum proses *training* pada klasifikasi, variabel input yang digunakan perlu dipilih untuk mendapatkan pola yang lebih baik, sehingga dimensionalitas data dapat dikurangi.

Pemilihan variabel yang dimaksudkan adalah memilih beberapa variabel yang ada, sehingga tidak semua variabel dalam *dataset* digunakan. Pada tugas akhir ini dianalisis penggunaan *Rough Set Approach* sebagai kriteria pemilihan variabel yang akan digunakan dalam klasifikasi pada *data mining*.

Dari uraian diatas, dirumuskan masalah dalam tugas akhir ini adalah :

1. Bagaimana jika jumlah variabel *dataset* terlalu banyak (*resource* yang dibutuhkan sangat besar pada saat *learning*).
2. Pemilihan variabel digunakan sebelum proses *learning* untuk mengurangi dimensionalitas *dataset*.

*Rough Set Approach* merupakan teknik untuk mengurangi atau menghapus suatu variabel yang mungkin *redundant*, tanpa menghilangkan informasi penting untuk pembuatan *classifier*. Hal ini berdasarkan konsep himpunan *upper* dan *lower approximation*, *approximation space* dan himpunan model.

Dalam Tugas Akhir ini, yang dibahas adalah implementasi penggunaan *Rough Set Approach* sebagai kriteria *variable selection* dalam *task classification* pada *Data Mining*, dengan batasan masalahnya sebagai berikut:

1. *Dataset* yang digunakan adalah *dataset* yang bertipe data record.
2. *Dataset* yang digunakan sudah dalam bentuk data diskret dan tidak menangani data kontinu.
3. Tidak menangani *missing value* dan *data cleaning*.
4. *Dataset* yang akan digunakan sebagai data *training* dan data *test* sudah tersedia.
5. Implementasi hanya pada perangkat lunak pemilihan variabel dari *dataset*, sedangkan untuk pembentukan *classifier* menggunakan perangkat lunak yang telah ada yaitu Weka.

### 1.3 Tujuan

Tujuan dari pembuatan tugas akhir ini adalah sebagai berikut :

1. Merancang dan membangun perangkat lunak untuk pemilihan variabel yang akan digunakan pada *task classification* dengan menggunakan *Rough Set Approach*.
2. Membandingkan hasil pemilihan variabel dengan perangkat lunak yang sudah ada, yaitu Weka dan Clementine.
3. Membandingkan hasil pengukuran *classifier* sebelum dan sesudah dilakukan *variable selection* menggunakan *Rough Set Approach* pada *dataset* yang meliputi *precision*, *recall* dan *accuracy* yang dihitung dari *confusion matrix* yang dihasilkan oleh Weka.

## 1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan untuk menyelesaikan tugas akhir ini adalah:

1. Studi Literatur

Pada tahap ini akan dilakukan pendalaman materi, identifikasi masalah dan metodologi yang akan digunakan dalam pemecahan masalah. Mempelajari teori dasar proses pemilihan variabel secara umum, teori dasar *Rough Set Approach* serta penggunaannya pada pemilihan variabel, cara pemilihan *core* yang dapat mewakili *rule* dari suatu *dataset*, pengukuran *classifier* berdasarkan *confusion matrix* yang dihasilkan oleh *software* klasifikasi yang digunakan. Setelah melakukan studi literatur akan dilakukan pengujian apakah metode *Rough Set Approach* dapat digunakan sebagai proses pemilihan variabel pada *data mining*.

2. Perancangan

Pada tahap ini dilakukan perancangan sistem yang akan dibangun, yang meliputi analisis terhadap kebutuhan perangkat lunak baik itu input maupun output sistem yang dibangun. Setelah itu dibuat diagram aliran data yang berfungsi untuk memberikan pennggambaran aliran data dan cara kerja sistem. Kemudian dilakukan pengumpulan data.

3. Implementasi

Impementasi pada program untuk pemilihan variabel dari *dataset* dengan menggunakan *Rough Set Approach* dimana variabel hasil pemilihan variabel akan digunakan pada klasifikasi di Weka untuk membentuk model.

4. Pengujian

Pada tahap ini dilakukan pengujian terhadap sistem yang dibangun, serta melakukan perbaikan terhadap *bug* dan *error* yang ditemukan pada perangkat lunak yang dibangun. Setelah dilakukan pengujian terhadap sistem, maka dilakukan proses pemilihan variabel pada *dataset* yang ada menggunakan sistem yang dibangun dan hasilnya akan digunakan pada klasifikasi di Weka.

5. Analisis dan pembuatan laporan

Melakukan analisis terhadap hasil pemilihan variabel sistem yang dibangun juga untuk hasil pemilihan variabel menggunakan Weka dan Clementine. Dari hasil pemilihan variabel yang dibangun, Weka dan Clementine akan dianalisis model yang dihasilkan dengan menggunakan weka yang meliputi parameter yang telah ditentukan sebelumnya. Setelah itu dilakukan pembuatan laporan hasil analisis.