**BAB VI**

**DATA MINING**

1. **TUJUAN**

* Dapat memahami kecerdasan buatan mengenai *Data Mining*
* Mengenal berbagai jenis *data mining* dan dapat mengimplementasikan menggunakan MATLAB

1. **PENDAHULUAN**

**Pengertian *Data Mining***

Secara istilah, *Data mining* adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basisdata. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basisdata.

Data mining terutama digunakan untuk mencari pengetahuan yang terdapat dalam basis data yang besar sehingga sering disebut *Knowledge Discovery in Databases* (KDD). Proses pencarian pengetahuan ini menggunakan berbagai teknik-teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksikannya. Proses pencarian bersifat iteratif dan interaktif untuk menemukan pola atau model yang benar, baru, bermanfaat, dan dapat dimengerti. Dalam penerapannya, dataminig memerlukan berbagai perangkat lunak analisis data untuk menemukan pola dan relasi data agar dapat digunakan untuk membuat prediksi dengan akurat.

**Manfaat Data Mining**

Pemanfaatan *data mining* dapat dilihat dari dua sudut pandang, yaitu sudut pandang komersial dan sudut pandang keilmuan. Dari sudut pandang komersial, pemanfaatan *data mining* dapat digunakan dalam menangani meledaknya volume data. Bagaimana mana menyimpannya, mengestraknya serta memanfaaatkannya. Berbagai teknik komputasi dapat digunakan menghasilkan informasi yang dibutuhkan. Informasi yang dihasilkan menjadi asset untuk meningkatkan daya saing suatu institusi.

*Data mining* tidak hanya digunakan untuk menangani persoalan menumpuknya data/informasi dan bagaimana menyimpannya tanpa kehilangan informasi yang penting (*warehousing*). *Data mining* juga diperlukan untuk menyelesaikan permasalahan atau menjawab kebutuhan bisnis itu sendiri, misalnya:

* Bagaimana mengetahui hilangnya pelanggan karena pesaing
* Bagaimana mengetahui produk atau konsumen yang memiliki kesamaan karakteristik
* Bagaimana mengidentifikasi produk-produk yang terjual bersamaan dengan produk lain.
* Bagaimana memprediski tingkat penjualan
* Bagaimana menilai tingkat resiko dalam menentukan jumlah produksi suatu item.
* Bagaimana memprediksi prilaku bisnis di masa yang akan datang

Dari sudut pandang keilmuan, data mining dapat digunakan untuk meng*capture*, menganalisis serta menyimpan data yang bersifat *real-time* dan sangat besar, misalnya:

* *Remote* sensor yang ditempatkan pada suatu satelit
* *Telescope* yang digunakan untuk memindai langit
* Simulasi saintifik yang membangkitkan data dalam ukuran *terrabytes*

Data mining merupakan salah satu metode alternatif yang dapat digunakan untuk mengolah data mentah, ketika metode konvensional tidak fisibel untuk dilakukan karena besarnya volume data yang diolah. Hal ini dapat terjadi karena *data mining* memiliki kemampuan mereduksi data baik melalui teknik katalogisasi, klasifikasi maupun segementasi.

**Proses Data Mining**

Data mining sesunggunghnya merupakan salah satu rangkaian dari proses pencarian pengetahuan pada *database* (*Knowledge Discovery in Database*/KDD). KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data. KDD adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti. Serangkaian proses tersebut yang memiliki tahap sebagai berikut (Tan, 2004):

1. Pembersihan data dan integrasi data (*cleaning and integration*). Proses ini digunakan untuk membuang data yang tidak konsisten dan bersifat *noise* dari data yang terdapat di berbagai basisdata yang mungkin berbeda format maupun *platform* yang kemudian diintegrasikan dalam satu *database datawarehouse*.
2. Seleksi dan transformasi data (*selection and transformation*). Data yang terdapat dalam *database datawarehouse* kemudian direduksi dengan berbagai teknik. Proses reduksi diperlukan untuk mendapatkan hasil yang lebih akurat dan mengurangi waktu komputasi terutama untuk masalah dengan skala besar (*large scale problem*). Beberapa cara seleksi, antara lain:

* *Sampling*, adalah seleksi *subset* representatif dari populasi data yang besar.
* *Denoising*, adalah proses menghilangkan *noise* dari data yang akan

ditransformasikan

* *Feature extraction*, adalah proses membuka spesifikasi data yang signifikan

dalam konteks tertentu

Transformasi data diperlukan sebagai tahap *pre-procecing*, dimana data yang diolah siap untuk ditambang. Beberapa cara transformasi, antara lain :

* *Centering*, mengurangi setiap data dengan rata-rata dari setiap atribut yang ada.
* *Normalisation*, membagi setiap data yang dicentering dengan standar deviasi

dari atribut bersangkutan

* *Scaling*, mengubah data sehingga berada dalam skala tertentu



**Gambar.1.** Tahap *Knowledge discovery* pada *Data Mining*.

1. Penambangan data (*data mining*)

Data-data yang telah diseleksi dan ditransformasi ditambang dengan berbagai teknik. Proses *data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan fungsi-fungsi tertentu. Fungsi atau algoritma dalam *data mining* sangat bervariasi. Pemilihan fungsi atau algoritma yang tepat sangat bergantung pada tujuan dan proses pencarian pengetahuan secara keseluruhan.

1. Evaluasi pola dan presentasi pengetahuan

Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami oleh pengguna.

**Jenis Data Mining**

Dari permasalahan yang diselesaikan menggunakan *data mining*, model penyelesaian permasalahan yang dilakukan adalah dengan membangun model berdasarkan data yang sudah digali untuk diterapkan terhadap data yang lain. Secara umum terdapat dua jenis tipologi dari *data mining* :

1. Metode Prediksi

Metode ini bermaksud memprediksi nilai yang akan datang berdasarkan data yang sudah ada. Penerapan metode ini biasanya digunakan untuk klasifikasi, deteksi *anomaly*, regresi, dll.

1. Metode Deskriptif

Merupakan metode yang bermaksud membantu *user* agar dapat melihat pola yang berasal dari data.

1. **LANGKAH PRAKTIKUM**

***Data Mining* dengan *K-Nearest Neighbor***

*Data Mining* dengan *K-Nearest Neighbor* biasanya digunakan untuk klasifikasi data yang didasarkan pada tetangga terdekat dari data yang diinginkan (*data Testing*).

Beberapa aturan yang ada pada data mining menggunakan *K-Neirest Neighbor* adalah sebagai berikut :

1. Tersedia beberapa data (*x*,*y*) yang terklasifikasi minimal 2 kelas.
2. Diberikan *query point* yang akan menentukan kelas dari data yang akan di uji.
3. Menghitung sejumlah *k* tetangga terdekat untuk menentukan kelas dari *query* tersebut berdasarkan *majority vote* (hasil terbanyak).

Sedangkan dalam merancang *Data mining* menggunakan *K-Nearest Neighbor*, dilakukan menggunakan Algoritma KNN Sebagai Berikut :

* Tentukan *k*
* Hitung jarak data baru dengan semua *data training*.
* Hitung jarak yang digunakan menggunakan *Euclidean distance* yang digambarkan dengan persamaan berikut :

Keterangan :

*d*: *Euclidean distance*

*q*: data baru

*p*: data *training*

* Urutkan jarak tersebut dari yang terdekat.
* Periksa kelas *k* tetangga terdekat
* Kelas data baru = kelas mayoritas *k* tetangga terdekatnya.

Contoh penerapan *data mining* menggunakan *K-Nearest Neighbor* adalah sebagai berikut :

Sebuah toko grosir, akan membeli telur dari seorang tengkulak. Dari telur-telur tersebut di peroleh data berdasarkan berat dan lebar telur untuk kemudian dimasukkan kedalam beberapa klasifikasi telur. Data telur yang sudah dimiliki oleh toko tersebut :

**Tabel1.** Data Training untuk Klasifikasi Telur.

|  |  |  |
| --- | --- | --- |
| Berat (gram) | Lebar (cm) | Klasifikasi |
| 8 | 6 | 1 |
| 8 | 4 | 1 |
| 4 | 3 | 2 |
| 3 | 4 | 2 |
| 1 | 5 | 3 |

Dari data tersebut, tergolong klasifikasi manakah telur dengan berat : 2 gram dengan lebar 6 cm, dengan k = 3?

Langkah penyelesaian :

1. Hitung terlebih dahulu titik sampel dengan *data training* menggunakan formula *Euclidean distance*:

(5)

(4)

(1)

(2)

(3)

1. Dari data tersebut, urutkan hasilnya dan ambil nilai terdekat sebanyak k=3
2. Periksa kelas dari ketiga tetangga tersebut,

– Klass 2

- Klass 2

– Klass 3

1. Dari pemeriksaan kelas tetangga, hasilnya lebih mengarah pada klasifikasi 2, sehingga data baru tersebut masuk dalam kategori klasifikasi 2.

Pada Matlab, KNN bisa dilakukan menggunakan *command* :

*knnclassify*(*sample*, *training*, *group*) atau

*knnclassify*(*sample*, *training*, *group*, *k*)

*sample* : matriks yang setiap barisnya merupakan titik *query*

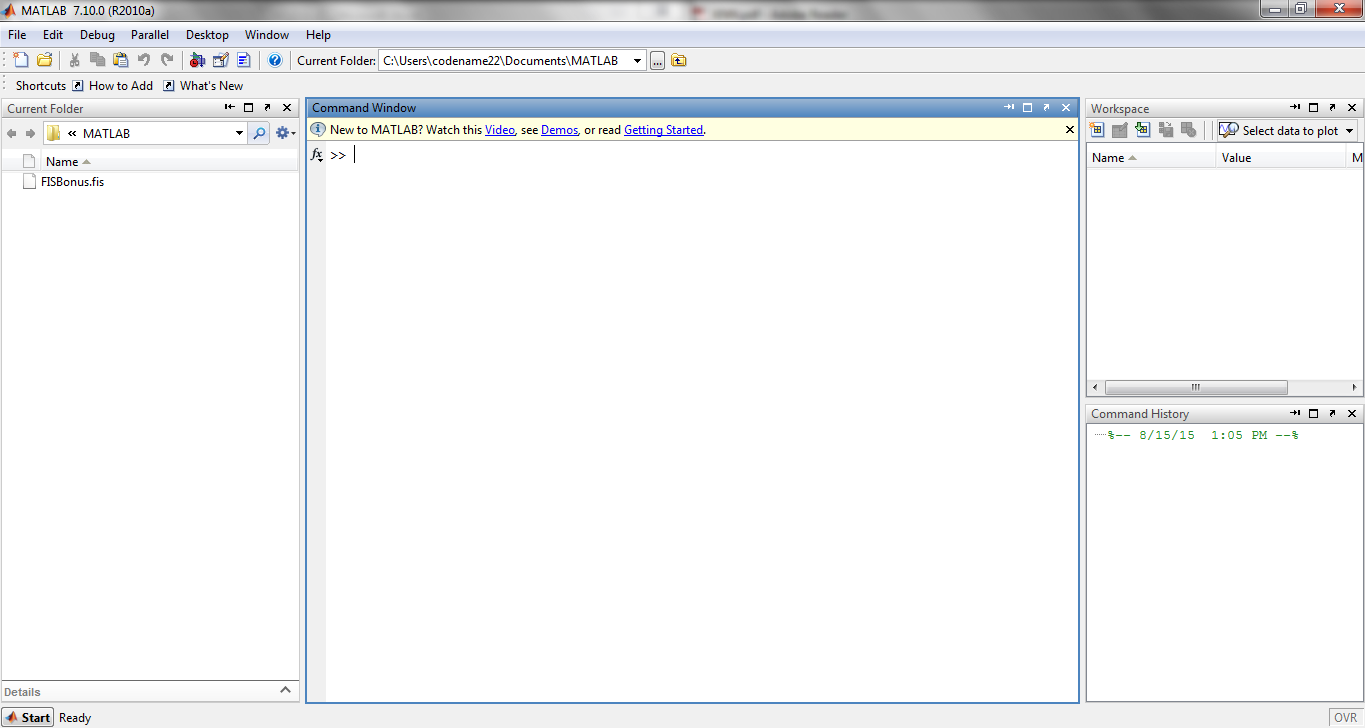
*training* : data yang akan diberikan sebagi calon *neighbor*

*group* : matrik hasil klasifikasi dari *training*

*k* : berapa *nearest neighbor* yang akan dihitung (*default* = 1)

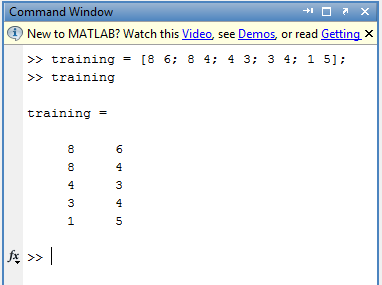
Langkah dalam membuat *K-Nearest Neighbour* menggunakan MATLAB

1. Buka halaman utama MATLAB



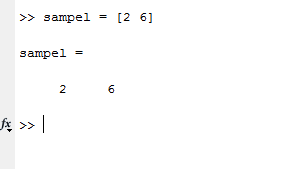
**Gambar 2.** Halaman Utama MATLAB

1. Buat *data training*nya terlebih dahulu dengan mengetikkan pada *command window*



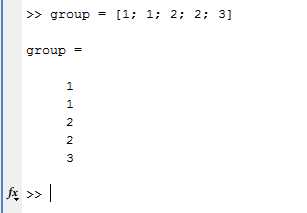
**Gambar 3.** Pengisian *Data Training* .

1. Setelah *data training* dibuat, berikutnya isikan *sampel* yang akan dihitung



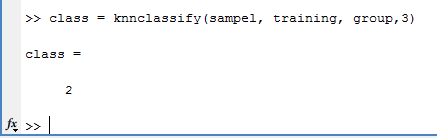
**Gambar 4.** Pengisian Data Sampel

1. Berikutnya buat nilai *group* untuk klasifikasi hasil dari *data training*



**Gambar 5.** Pengisian Nilai *Group* untuk *data training*.

1. Klasifikasikan data tersebut dengan perintah *knnclassify*.



**Gambar 6.** Hasil Klasifikasi dari data sampel berdasarkan *data training* menggunakan *K-Nearest Neighbor*

1. **TUGAS PRAKTIKUM**

Dalam studi kasus lulusan mahasiswa teknik informatika untuk menentukan mahasiswa dengan status kelulusannya yang berdasarkan pada besar IPK, lama studi dan jumlah keikutsertaan organisasi. Dari data yang sudah ada pada program studi adalah sebagai berikut :

|  |  |  |  |
| --- | --- | --- | --- |
| IPK | Lama Studi | Jumlah Organisasi | Kelulusan |
| 3.81 | 3.5 | 5 | 4 |
| 3 | 7 | 1 | 2 |
| 3.67 | 4.1 | 3 | 4 |
| 3.32 | 3.8 | 5 | 3 |
| 3 | 5 | 2 | 2 |
| 3.20 | 4 | 4 | 3 |
| 2.75 | 6 | 3 | 1 |
| 3 | 5 | 5 | 2 |
| 2.52 | 7 | 0 | 1 |
| 3.47 | 4.6 | 3 | 3 |
| 3.9 | 4.2 | 2 | 4 |

Keterangan : 4 : Istimewa, 3: Memuaskan , 2 : Baik, 1 : Cukup

Dari data tersebut, jika ada lulusan dengan ipk: 3.15 dengan lama studi 3.5 th dan jumlah organisasi 5, maka tergolong mahasiswa dengan status kelulusan yang mana?