PEMBUATAN APLIKASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORHOOD

(Studi Kasus Data Akademik Jurusan Teknik Komputer-S1 Universitas Komputer Indonesia)

Oleh:

Astrid Darmawan

10207104

Pembimbing:

Selvia Lorena Br. Ginting, M.T

Wendi Zarman, M.Si

JURUSAN TEKNIK KOMPUTER

FAKULTAS TEKNIK DAN ILMU KOMPUTER

UNIVERSITAS KOMPUTER INDONESIA

2012

Pendahuluan

- Latar Belakang
- Maksud dan Tujuan
- Batasan Masalah

Latar Belakang

- Adanya tumpukan data

Maksud:

- Membuat perancangan aplikasi untuk mengklasifikasi masa studi mahasiswa di Jurusan Teknik Komputer dengan menggunakan Algoritma K-Nearest Neighborhood.

Tujuan:

- Untuk memprediksi tingkat kelulusan dan persentase kelulusan mahasiswa di Jurusan Teknik Komputer khususnya S1.

Batasan Masalah

- Data yang digunakan hanya data Indeks Prestasi (IP) mahasiswa jurusan Teknik Komputer Program Sarjana (S1) dari semester 1 sampai semester 6.
- Data yang digunakan telah dibersihkan secara manual sebelum diproses *mining*.
- Visual Basic 6.0 sebagai aplikasi sistem antarmuka.
- Pengolahan basis data menggunakan *database* Microsoft Access 2007.
- Data yang digunakan adalah data akademik mahasiswa Jurusan Teknik Komputer tahun 2001-2006

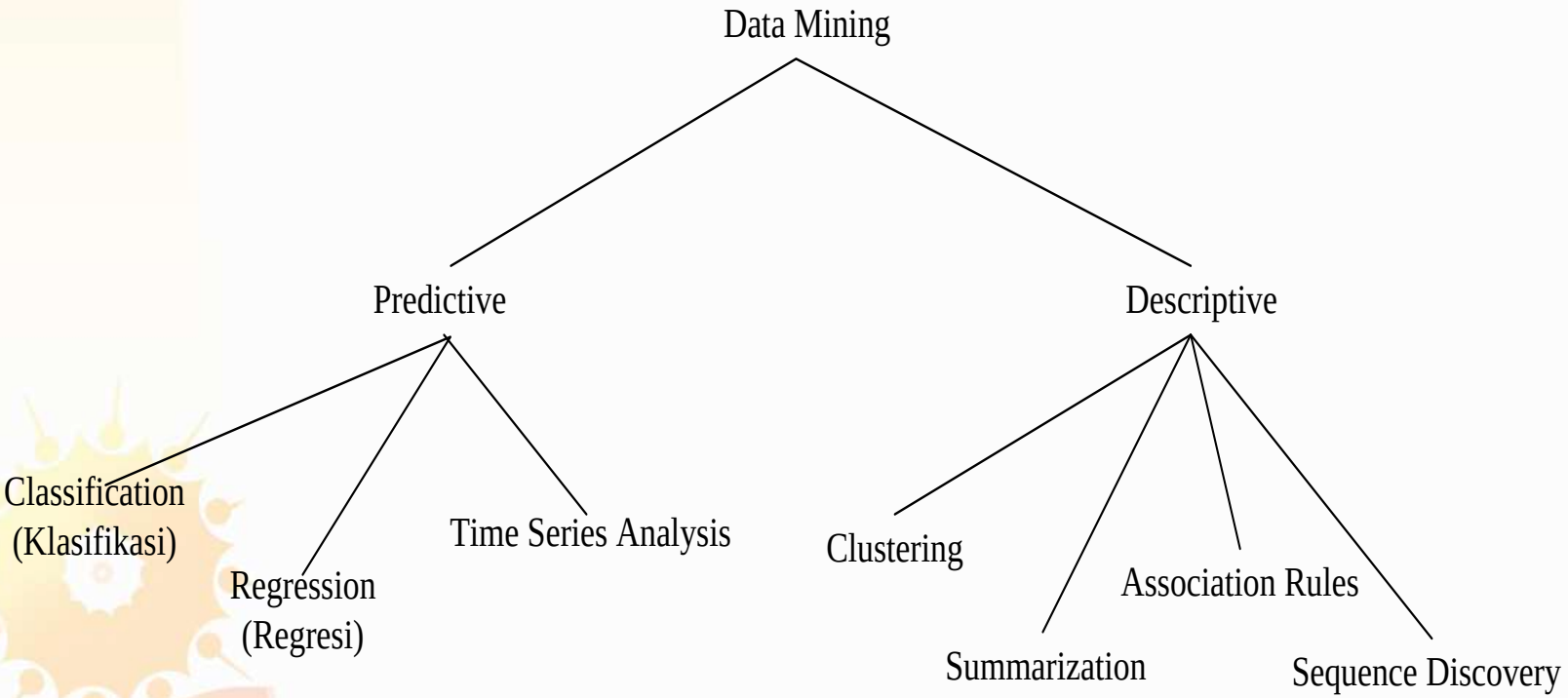
Dasar Teori

- Data Mining
- Klasifikasi
- Algoritma Nearest Neighborhood

Data Mining

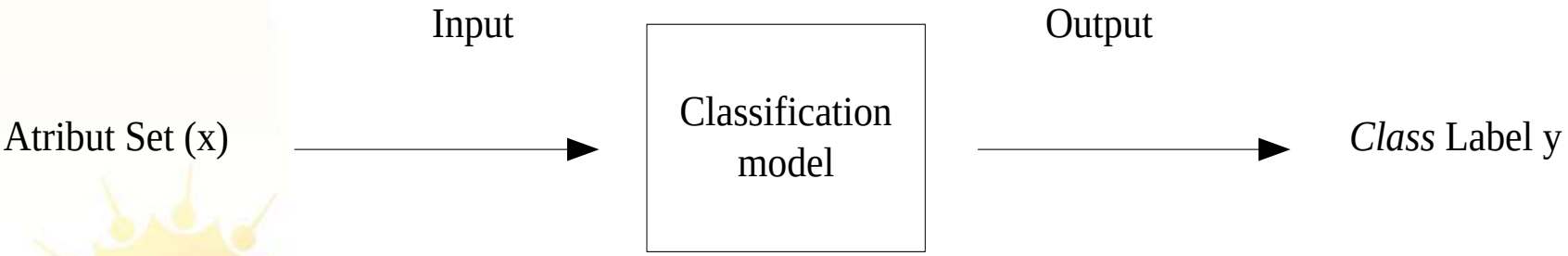
Data mining merupakan proses menemukan pengetahuan yang baru dari data yang berjumlah besar yang disimpan di dalam *database*, gudang data atau repositori informasi.

Pengelompokan Data Mining



Klasifikasi

Klasifikasi: tugas pembelajaran sebuah fungsi target f yang memetakan setiap himpunan atribut x ke salah satu *class label* y yang telah didefinisikan sebelumnya.



Gambar.2.2 Model Klasifikasi

Algoritma Nearest Neighborhood (NN)

Nearest-neighbor

Pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru (*testing data*) dengan kasus lama (*training sample*).

Jenis algoritma *Nearest-neighbor:*

1-NN, yaitu pengklasifikasikan dilakukan terhadap 1 *labeled data* terdekat.

K-NN, yaitu pengklasifikasikan dilakukan terhadap k *labeled data* terdekat dengan K>1.

Algoritma K-Nearest Neighborhood (KNN)

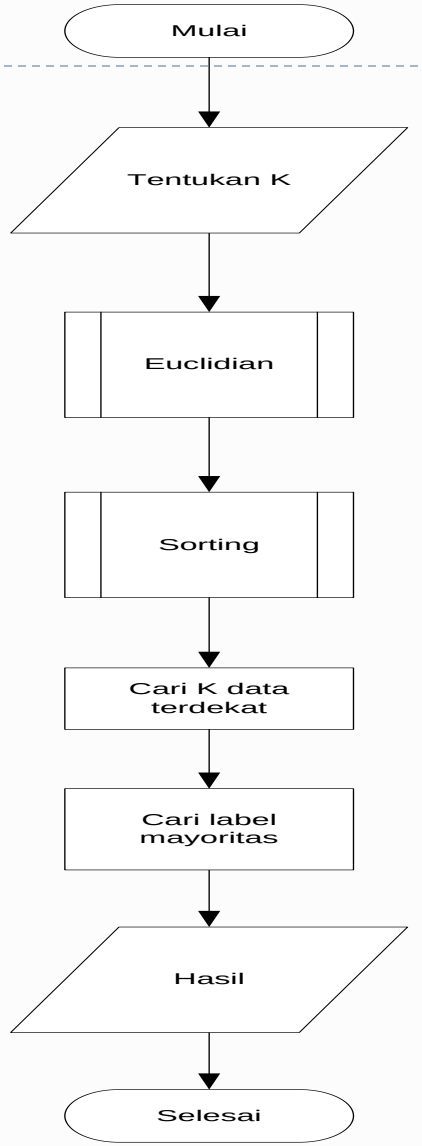
Ecludian Distance:

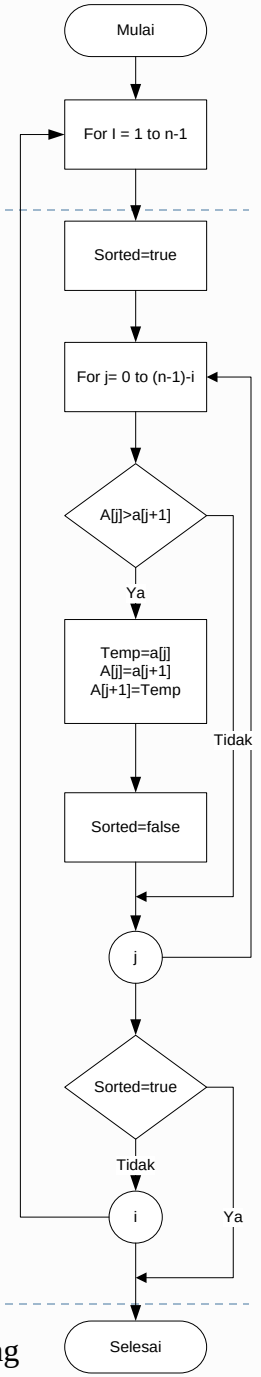
$$\left(\sum_{k=1}^{m}\left(x_{ki}-x_{kj}\right)^2\right)^{1/2}$$

Atribut Data Training

Atribut Data Testing

Persamaan 2.1





Gambar 3.13 (A) Algoritma KNN & (B) Algoritma Sorting

Perancangan

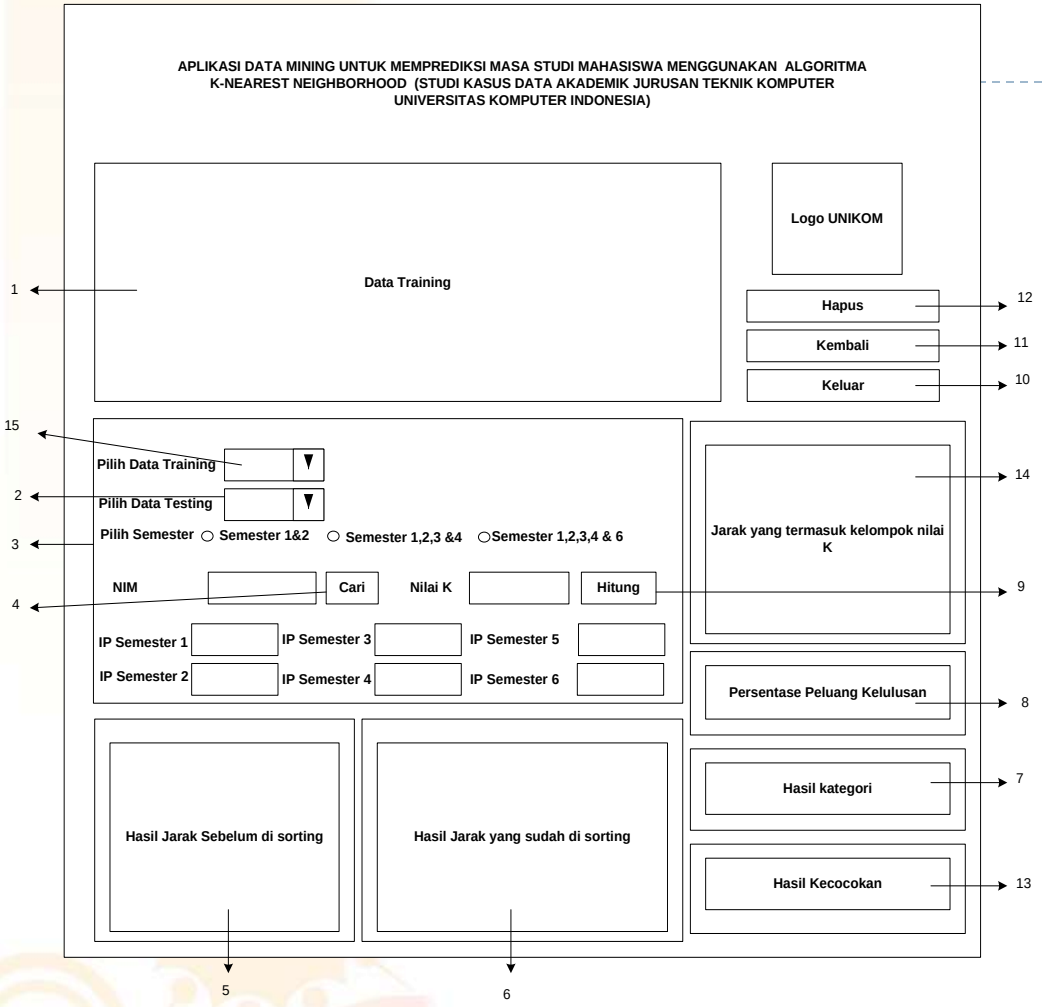
Aplikasi data mining yang dibuat terdiri dari dua data, yaitu:

- Data Testing
  1. NIM
  2. Indeks Prestasi (IP) mahasiswa dari semester satu sampai enam.
- Data Training
  1. NIM
  2. Indeks Prestasi (IP) mahasiswa dari semester satu sampai enam.
  3. Keterangan (Kategori Kelulusan)

Data training memiliki kategori sebagai berikut:

Tabel 3.4 Kategori Kelulusan Berdasarkan Lama Studi

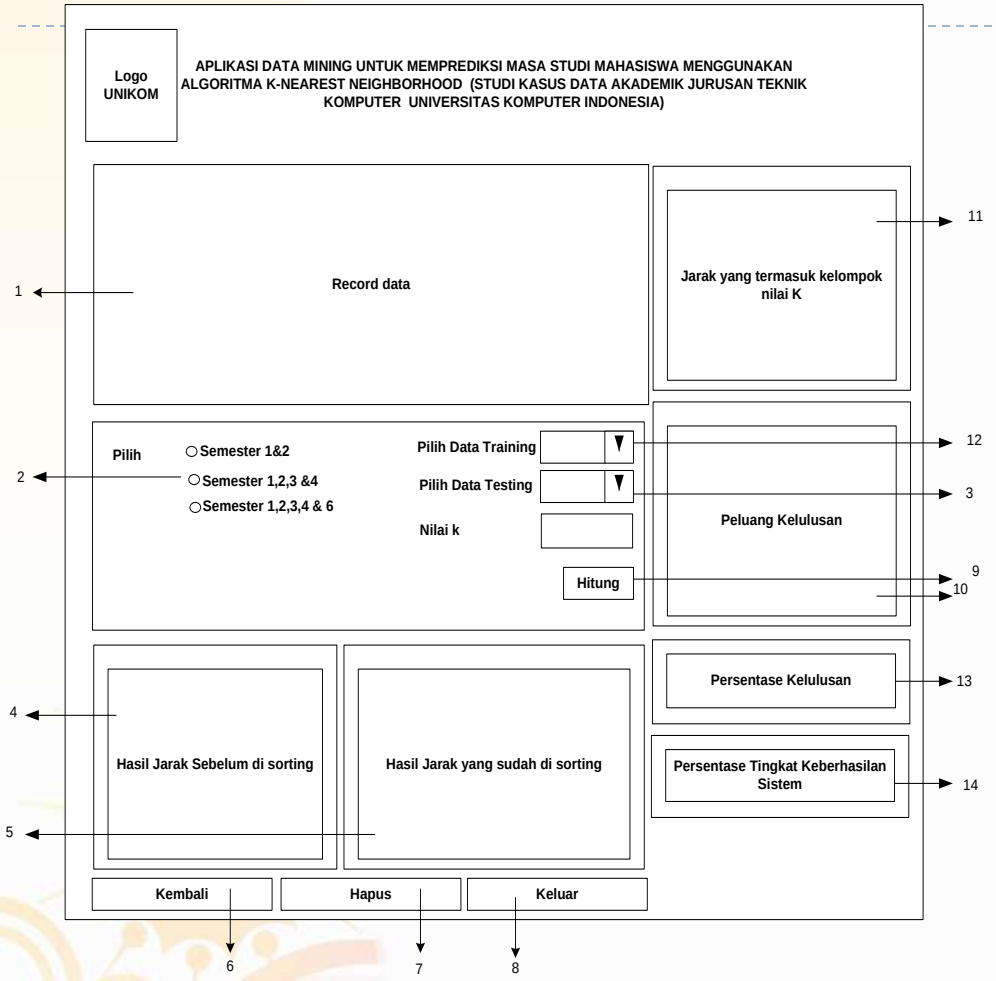
Lulus	Kategori
Lulus <= 5 Tahun	Ya
Lulus > 5 Tahun	Tidak



Perancangan Aplikasi Data Mining untuk 1 Data Mahasiswa

Tabel 3.17 Tabel Penjelasan tentang PerancanganForm Orang

No	Keterangan
1	Data Grid untuk menampilkan data *training*.
2	Pemilihan Tabel*database data testing*
3	Option untuk memilih semester.
4	*Button* untuk mencari NIM.
5	Listbox untuk keluaran hasil perhitungan jarak sebelum di sorting.
6	Listbox untuk keluaran hasil perhitungan jarak setelah di sorting.
7	Listbox untuk kategori
8	Listbox untuk peluang kelulusan
9	*Button* untuk hitung jarak.
10	*Button* untuk keluar
11	*Button* untuk menghapus.
12	*Button* untuk kembali ke Form Utama.
13	Listbox untuk hasil kecocokan hasil data uji tehadap data asli.
14	Listbox untuk hasil dari nilai K
15	Pemilihan Tabel*database data training*



Tabel 3.20 Tabel Penjelasan tentang Form Keseluruhan

No	Keterangan
1	Data Grid untuk menampilkan *datatraining*dan *data testing.*
2	Option untuk pemilihan semester .
3	combobox untuk memilih Tabel data *testing.*
4	List box untuk hasil jarak baru sebelum disorting
5	List box untuk menampilkan hasil jarak yang sudah disorting.
6	*Button* untuk kembali ke form utama.
7	*Button* untuk menghapus.
8	*Button* untuk keluar dari aplikasi *data mining.*
9	Buttin untuk menghitung jarak
10	Listbox untuk keluaran hasil klasifikasi.
11	Listbox untuk hasil pengelompokan nilai K.
12	combobox untuk memilih Tabel data *training.*
13	Listbox untuk menampilkan persentase kecocokan data.
14	Listbox untuk menampilkan persentase prediksi kelulusan.

Perancangan Aplikasi Data Mining untuk 1 *Database*

Contoh Kasus

Tabel .1 data *training*

M	SEMESTER						Y=Klasifikasi
	X_{11} = 1	X_{12}=2	X_{13}=3	X_{14}=4	X_{15}=5	X_{16}=6	
1	3.94	4	4	4	3.89	3.58	Ya
2	1.68	1.84	1.53	2.32	2.88	2.69	Tidak
3	2.63	1.63	2.42	2.58	3	3.56	Tidak
4	3.21	3.42	3.84	2.79	3.6	3.75	Ya
5	3.11	3.58	3.35	3	3.11	3.21	Ya
6	2.68	2.79	3.16	2.95	3.23	3.68	Ya
7	2	1.53	2.11	2.16	2	1.88	Tidak

Label class y

Tabel 3.6 Data *Testing*

M	SEMESTER						Y=Klasifikasi
	X_{21} = 1	X_{22}=2	X_{23}=3	X_{24}=4	X_{25}=5	X_{26}=6	
1	2.95	2.76	2.32	1.8	2.75	2.87	?

Label class y

Proses

- Parameter yang dipakai adalah K=5
- Menghitung kuadrat jarak *Euclid* (*query instance*) masing-masing objek terhadap sampel data atau *training sample* yang diberikan dengan menggunakan rumus:

$$\left(\sum_{k=1}^{m}\left(x_{ik}-x_{ij}\right)^2\right)^{1/2}$$

- M=1

Query instance

$$= ((x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2$$
$$+ (x_{14} - x_{24})^2 + (x_{15} - x_{25})^2 + (x_{16} - x_{26})^2)^{1/2}$$
$$= ((3.94 - 2.95)^2 + (4 - 2.76)^2 + (4 - 2.32)^2$$
$$+ (4 - 1.8)^2 + (3.89 - 2.75)^2$$
$$+ (3.58 - 2.87)^2)^{1/2}$$
$$= ((0.99)^2 + (1.24)^2 + (1.68)^2 + (2.2)^2 + (1.14)^2$$
$$+ (0.71)^2)^{1/2}$$
$$= (0.98 + 1.54 + 2.82 + 4.48 + 1.30 + 0.50)^{1/2}$$
$$= (11.98)^{1/2}$$
$$= 3.46$$

dst sampai M=7

Tabel 3.7 *Square Instance to Query Distance*

M	SEMESTER						Square instance to query distance
	X_{11} =1	X_{12}=2	X_{13}=3	X_{14}=4	X_{15}=5	X_{16}=6	
1	3.94	4	4	4	3.89	3.58	3.46
2	1.68	1.84	1.53	2.32	2.88	2.69	1.84
3	2.63	1.63	2.42	2.58	3	3.56	1.59
4	3.21	3.42	3.84	2.79	3.6	3.75	2.3
5	3.11	3.58	3.35	3	3.11	3.21	1.86
6	2.68	2.79	3.16	2.95	3.23	3.68	1.73
7	2	1.53	2.11	2.16	2	1.88	2.03

- Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak *Euclid* terkecil.

Tabel 3.8 Mengurutkan Objek ke dalam Kelompok ke Jarak Euclid Terkecil

m	SEMESTER						Square instance to query distance	Jarak terkecil	Apakah termasuk nearest-neighbor (k)
	X_{11} = 1	X_{12}=2	X_{13}=3	X_{14}=4	X_{15}=5	X_{16}=6			
1	3.94	4	4	4	3.89	3.58	3.46	7	-
2	1.68	1.84	1.53	2.32	2.88	2.69	1.84	3	Ya
3	2.63	1.63	2.42	2.58	3	3.56	1.59	1	Ya
4	3.21	3.42	3.84	2.79	3.6	3.75	2.3	6	-
5	3.11	3.58	3.35	3	3.11	3.21	1.86	4	Ya
6	2.68	2.79	3.16	2.95	3.23	3.68	1.73	2	Ya
7	2	1.53	2.11	2.16	2	1.88	2.03	5	Ya

- Mengumpulkan label *class* Y (klasifikasi *nearest neighbor*).

Tabel 3.9 Label *Class Y*

M	SEMESTER						Square instance to query distance	Jarak terkecil	Apakah termasuk nearest-neighbor (k)	Y=Klasifikasi
	X_{11} =1	X_{12}=2	X_{13}=3	X_{14}=4	X_{15}=5	X_{16}=6				
1	3.94	4	4	4	3.89	3.58	3.46	7	-	Ya
2	1.68	1.84	1.53	2.32	2.88	2.69	1.84	3	Ya	Tidak
3	2.63	1.63	2.42	2.58	3	3.56	1.59	1	Ya	Tidak
4	3.21	3.42	3.84	2.79	3.6	3.75	2.3	6	-	Ya
5	3.11	3.58	3.35	3	3.11	3.21	1.86	4	Ya	Ya
6	2.68	2.79	3.16	2.95	3.23	3.68	1.73	2	Ya	Ya
7	2	1.53	2.11	2.16	2	1.88	2.03	5	Ya	Tidak

Mencari Mayoritas Kategori

Tabel 6 Hasil Akhir

m	SEMESTER						Square instance to query distance	Jarak terkecil	Apakah termasuk nearest-neighbor (k)	Y=Klasifikasi
	X_{11} =1	X_{12}=2	X_{13}=3	X_{14}=4	X_{15}=5	X_{16}=6				
1	3.94	4	4	4	3.89	3.58	3.46	7	-	Ya
2	1.68	1.84	1.53	2.32	2.88	2.69	1.84	3	Ya	Tidak
3	2.63	1.63	2.42	2.58	3	3.56	1.59	1	Ya	Tidak
4	3.21	3.42	3.84	2.79	3.6	3.75	2.3	6	-	Ya
5	3.11	3.58	3.35	3	3.11	3.21	1.86	4	Ya	Ya
6	2.68	2.79	3.16	2.95	3.23	3.68	1.73	2	Ya	Ya
7	2	1.53	2.11	2.16	2	1.88	2.03	5	Ya	Tidak

Ya=2

Tidak=3

Jadi, Data Testing tersebut termasuk kategori Lulus > 5

PENGUJIAN

- Pengujian terdiri dari 2 proses yaitu:
  1. Pengujian 1 *database* mahasiswa (data *training = 30*)
  2. Pengujian 1 *database* mahasiswa (data *training = 61*)

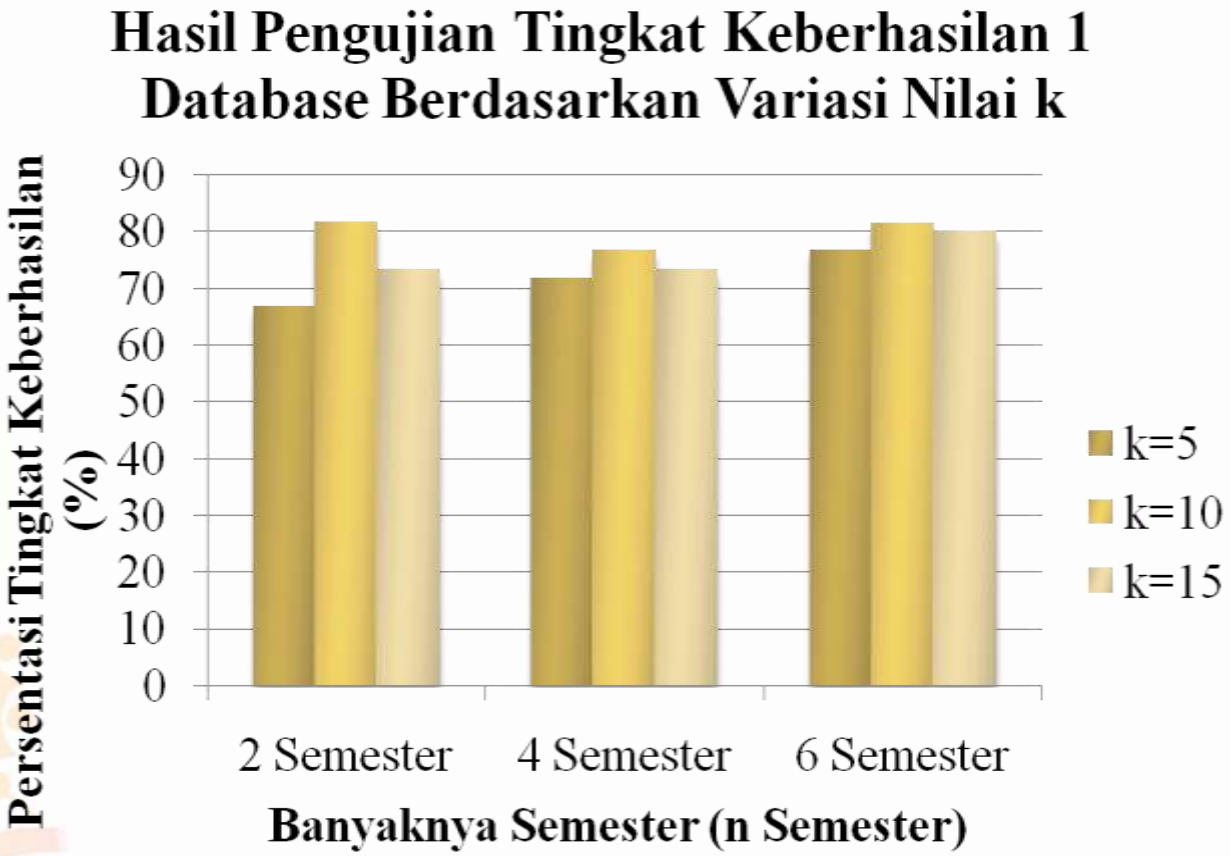
Masing-masing proses pengujian tersebut menggunakan Indeks Prestasi (IP)

1. Dua semester (semester 1 dan 2),
2. Empat semester (semester 1-4) dan
3. Enam semester (semester 1-6)

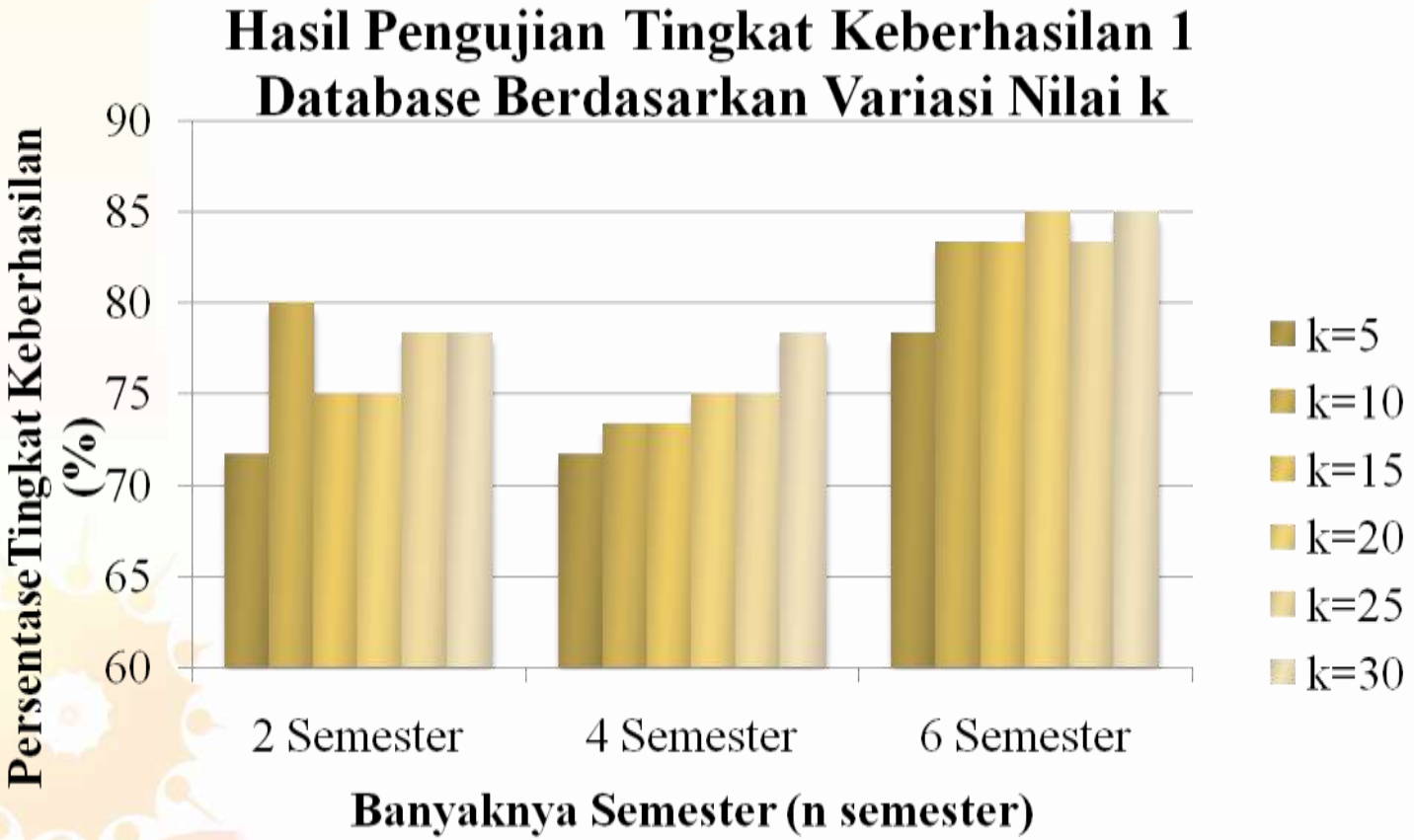
Menggunakan nilai k yang berbeda.

Untuk mengetahui tingkat keberhasilan pada sistem ini maka penulis menggunakan rumus:

$$Tingkat\ Keberhasilan\ Sistem = \frac{\sum hasil\ pengujian\ bernilai\ benar\ (kecocokan)}{\sum banyaknya\ data\ sampel}\ x\ 100\%$$



Gambar 4.12 *Testing* dengan 1 *Database* mahasiswa (*data training*=30 data)



Gambar 4.19 *Testing* dengan 1 *Database* Mahasiswa (*data training*=61 data)

Kesimpulan

1. Aplikasi *data mining* ini telah berhasil membuat aplikasi yang dapat memprediksi masa studi mahasiswa menggunakan algoritma *k-nearest neighborhood* yang akan diterapkan di Jurusan Teknik Komputer.
2. Aplikasi *data mining* ini dapat memprediksi dengan menggunakan 1 data mahaisawa atau 1 *database* mahasiswa.
3. Untuk menggunakan *data training* yang berjumlah 30 data dengan menguji *data testing* berjumlah 60 data, maka didapatkan nilai k yang terbaik untuk memprediksi masa studi mahasiswa yaitu sebagai berikut:
   - Untuk dua semester yaitu nilai k yang terbaik untuk digunakan memprediksi studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 81.66%.
   - Untuk empat semester yaitu nilai k yang terbaik untuk digunakan memprediksi studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 76.66%.
   - Untuk enam semester yaitu nilai k=10merupakan nilai k yang terbaik untuk digunakan memprediksi masa studi mahasiswa dengan tingkat keberhasilan 81.66%.
4. Untuk menggunakan *data training* yang berjumlah 30 data dengan menguji *data testing* berjumlah 60 data, maka didapatkan nilai k yang terbaik untuk memprediksi masa studi mahasiswa yaitu sebagai berikut:
   - Untuk dua semester yaitu nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 80%.
   - Untuk empat semester, nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=30 dengan tingkat keberhasilan 78.33%.
   - Untuk enam semester, nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=20 dan k=30 dengan tingkat keberhasilan 85%.
5. Setelah melakukan pengujian dengan melakukan perubahan pada nilai k, maka akan menghasilkan prediksi kelulusan yang bervariasi.
6. Nilai k yang terbaik juga tergantung pada jumlah data yang digunakan. Ukuran nilai k yang besar untukmemprediksi masa studi mahasiswa belum tentu menjadi nilai k yang terbaik dengan tingkat keberhasilan yang tinggi begitupun juga sebaliknya.

SARAN

- Aplikasi *data mining* ini dapat dibuat dengan menggunakan jaringan LAN sehingga nanti tidak hanya digunakan oleh satu *pengguna* di satu tempat saja tetapi dapat digunakan oleh banyak pengguna (para dosen dan pihak jurusan) di tempat yang berbeda.
- Aplikasi *data mining* ini belum memperhitungkan faktor luar yang terjadi pada mahasiswa tersebut sehingga mahasiswa tersebut tidak lulus tepat waktu. Faktor luar tersebut dapat menjadi sebuah parameter baru dalam aplikasi ini.
- Kasus ini dapat diterapkan dengan menggunakan metode algoritma *data mining* yang lainnya, misalnya ID3.
- Penulis juga menyarankan agar aplikasi *data mining* ini dapat diterapkan menggunakan teknik data *warehouse*, sehingga tidak hanya menggunakan data akademik 1 jurusan saja tetapi dapat menggunakan banyak data akademik dari jurusan lain.

Terima Kasih