# BAB 2

# TINJAUAN PUSTAKA DAN DASAR TEORI

## 2.1 Tinjauan Pustaka

Aplikasi *Data Mining* untuk memprediksi masa studi mahasiswa menggunakan Algoritma *K-Nearest Neighborhood* yang dibuat oleh Darmawan(2012) dengan menggunakan atribut nilai semester 1 sampai dengan 6. Prediksi dilakukan pada semester 2, 4, dan 6 dengan *data testing* sebanyak 60 dan *data training* sebanyak 30 diperoleh akurasi sebesar 81,66% dengan nilai k = 10.

Aplikasi berbasis Web untuk memprediksi Masa Studi Mahasiswa menggunakan *Data Mining* dengan metode algoritma K- Nearest Neighbor oleh Setyawan(2014). Pada penelitian ini kriteria yang digunakan adalah Nilai Rata-rata Ebtanas Murni (NEM), Nilai rata-rata STTB, Propinsi Asal SMA, Status SMA (Negeri/Swasta), jurusan SMA(IPA/IPS) dan Jenis Kelamin. Hasil pengujian menggunakan proses prediksi diketahui bahwa calon mahasiswa dengan NEM, STTB yang tinggi tidak berpengaruh terhadap masa studi dan tinggi rendahnya IPK.

Model Algoritma K-Nearest Neighbor(K-NN) untuk Prediksi Kelulusan Mahasiswa dibuat oleh Rohman (2015). Pada penelitian ini menggunakan atribut umur, Indeks Prestasi Semester 1 sampai dengan 8. Yang dihasilkan dari penelitian ini adalah nmendapatkan nilai akurasi dan AUC dari algoritma klasifikasi *data mining* dengan menggunakan algoritma K-NN. Dalam penelitian

ini dalam memprediksi kelulusan mahasiswa dengan menggunakan algoritma klasifikasi data mining *K-Nearest Neighbor* dengan mengklaster *data k=1, k=2, k=3, k=4*, dan *k=5*. Hasil yang diperoleh dengan cluster data *k=5 accuracy* adalah 85,15% dan nilai *AUC* adalah 0.888 adalah akurasi paling tinggi.

Aplikasi berbasis Web untuk memprediksi Masa Studi Mahasiswa menggunakan Data Mining dengan metode algoritma K- Nearest Neighbor yang dibuat oleh Mustafa dan Simpen (2015) . Atribut yang digunakan adalah nilai ujian nasional (UN), asal sekolah/ daerah, jenis kelamin, pekerjaan dan penghasilan orang tua, jumlah bersaudara. Dari hasil pengujian dengan menerapkan algoritma KNN dan menggunakan data sampel alumni tahun wisuda 2004 s.d. 2010 untuk kasus lama dan data alumni tahun wisuda 2011 untuk kasus baru diperoleh tingkat akurasi sebesar 83,36%.

Selanjutnya adalah skripsi yang dibuat yaitu untuk memperkirakan masa studi mahasiswa dengan menggunakan metode *K- Nearest Neighbor* dengan menggunakan atribut Indeks Prestasi Semester (IPS) 1 sampai dengan 4 dan jumlah SKS yang ditempuh pada semester 4. Perbedaan dengan penelitian sebelumnya adalah lokasi penelitian yang akan dikerjakan dan atribut yang digunakan yaitu ips semester 1 sampai dengan 4. Alasan pemilihan atribut yaitu dengan memprediksi mahasiswa lebih awal yaitu pada semester 4 supaya dapat melakukan perubahan prestasi akademik mahasiswa untuk mencapai target kelulusan yang diinginkan. Tabel perbandingan dari penelitian sebelumnya dapat dilihat pada tabel 2.1.

**Tabel 2.1 Perbandingan Penelitian Sebelumnya**

| No | Penulis | Objek Penelitian | Metode | Atribut |
|---|---|---|---|---|
| 1 | Darmawan(2012 ) | Jurusan Teknik Komputer-S1, Universitas Komputer Indonesia | K-NN | - Indeks Prestasi Semester 1, 2, 3, 4, 5, 6 |
| 2 | Rohman (2014) | Universitas Pandanaran Semarang | K-NN | - Jenis Kelamin<br>- Umur<br>- Indeks Prestasi semester 1 sampai dengan 8 |
| 3 | Setyawan (2014) | Mahasiswa Teknik Informatika Fakultas Teknik Industri UPN"Veteran" Yogyakarta | K- NN | - Nilai Rata- rata Ebtanas Murni (NEM)<br>- Nilai rata-rata STTB<br>- Propinsi Asal SMA<br>- Status SMA (Negeri/Swasta)<br>- jurusan SMA(IPA/IPS)<br>- Jenis Kelamin. |
| 4 | Mustafa dan Simpen (2015) | Akademik Mahasiswa STMIK Dipanegara Makassar | K- NN | - Jenis kelamin<br>- Agama<br>- Nem<br>- Jurusan<br>- Provinsi SMA |
| 5 | Skripsi(2017) | Program Studi Sistem Informasi STMIK Akakom Yogyakarta | K-NN | - Indeks Prestasi Semester 1, 2, 3, 4<br>- Jumah SKS semester 4 |

## 2.2 Dasar Teori

### 2.2.1 Data Mining

Data mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan. Data mining ini juga dikenal dengan istilah *pattern recognition* (Santoso, 2007).

Data mining merupakan metode pengolahan data berskala besar oleh karena itu data mining ini memiliki peranan penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi. Secara umum kajian data mining membahas metode-metode seperti, *clustering*, klasifikasi, regresi, seleksi variable, dan market basket analisis (Santoso, 2007).

Proses dan teknik penyaringan data menentukan mutu pengetahuan dan informasi yang akan diperoleh. Istilah lain untuk data mining adalah *Knowledge Discovery in Databases* (KDD). KDD merupakan sebuah proses yang terdiri dari serangkaian proses interasi yang terurut, dan data mining merupakan salah satu langkah dalam proses KDD Urutan langkah dalam KDD adalah sebagai berikut:

1. ***Data Selection***

Pemilihan (seleksi) data dari sekumpulan data operasi perlu dilakukaan sebelum tahap penggalian informasi dalam KDD dimulai.

2. ***Pre-processing/ Cleaning***

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membangun duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*).

3. ***Transformation***

*Coding* adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangan tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

***4. Data Mining***

*Data mining* dalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam data mining sangat bervariasi.

***5. Interpretation/Evaluation***

Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

### 2.2.2 K- Nearest Neighbor (K-NN)

Algoritma Nearest Neighbor adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. (Kusrini dan Luthfi, 2009)

Algoritma K-NN adalah suatu metode yang menggunakan algoritma *supervised.* Perbedaan antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada *unsupervised learning*, data belum memiliki pola apapun, dan tujuan *unsupervised learning* untuk menemukan pola dalam sebuah data. Tujuan

dari algoritma *k*-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples*.

*K-Nearest Neighbor* sering digunakan dalam klasifikasi dengan tujuan dari algoritma ini adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples.* Algoritma *K-Nearest Neighbor* (K-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Teknik ini sangat sederhana dan mudah diimplementasikan. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Sebuah titik pada ruang ini ditandai kelas *c* jika kelas *c* merupakan klasifikasi yang paling banyak ditemui pada *k* buah tetangga terdekat titik tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak *Euclidean*.

Untuk mendefinisikan jarak antara dua titik yaitu titik pada *data testing* (x) dan titik pada *data training* (y) maka digunakan rumus *Euclidean distance* (Santoso, 2007).

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2.1)$$

atau

$$d(x,y) = \begin{matrix} y_1 \\ x_1 - ¿ \\ ¿ \\ x \\ ¿ \\ y_2 \\ ¿ 2 - ¿ \\ ¿ \\ x \\ ¿ \\ y_n \\ ¿ n - ¿ \\ ¿ \\ ¿ \\ \sqrt{¿} \end{matrix} \qquad (2.2)$$

Keterangan:

x = *Data Testing*

y = *Data Training*

d = Jarak

n, i = Dimensi data

Sedangkan untuk menghitung akurasi data mengunakan rumus seperti dibawah ini:

$$\text{Akurasi} = \frac{jumlah\ pengujian\ yang\ diprediksi\ benar}{Jumlah\ data\ yang\ diuji} x\ 100$$

(2.3)

$$\text{Persentase erorr} = \frac{jumlah\ pengujian\ yang\ diprediksi\ salah}{Jumlah\ data\ yang\ diuji} x\ 100$$

(2.4)

Berikut ini merupakan contoh perhitungan menggunakan metode K – NN dengan mengambil data training sebanyak 20 data dari alumni :

**Tabel 2.2 Data training**

| NO | NIM | IPS 1 | IPS 2 | IPS 3 | IPS 4 | TOTAL SKS | STATUS |
|---|---|---|---|---|---|---|---|
| 1 | 135610075 | 3,19 | 3,52 | 3,50 | 3,13 | 91 | LC |
| 2 | 135610082 | 3,19 | 3,52 | 3,33 | 3,22 | 91 | LC |
| 3 | 135610123 | 3,86 | 3,65 | 3,36 | 3,68 | 88 | LC |
| 4 | 135610140 | 3,38 | 3,38 | 3,29 | 3,33 | 93 | LC |
| 5 | 135610152 | 2,62 | 2,86 | 3,60 | 3,42 | 86 | LC |
| 6 | 135610157 | 3,62 | 3,38 | 3,26 | 3,67 | 89 | LC |
| 7 | 085610034 | 1,00 | 1,38 | 0,00 | 1,13 | 70 | LL |
| 8 | 095610111 | 2,14 | 0,79 | 1,62 | 0,13 | 68 | LL |
| 9 | 095610116 | 1,10 | 1,55 | 0,00 | 2,20 | 63 | LL |
| 10 | 105610080 | 1,48 | 2,62 | 2,29 | 1,54 | 70 | LL |
| 11 | 105610087 | 2,24 | 0,00 | 1,92 | 2,57 | 74 | LL |
| 12 | 115610022 | 1,76 | 1,93 | 2,43 | 1,90 | 70 | LL |
| 13 | 115610062 | 2,05 | 2,19 | 1,59 | 1,15 | 82 | LL |
| 14 | 115610118 | 2,14 | 2,50 | 2,80 | 2,33 | 82 | LL |
| 15 | 125610042 | 2,10 | 2,24 | 2,50 | 2,52 | 81 | LT |
| 16 | 125610050 | 2,14 | 2,56 | 2,90 | 2,67 | 83 | LT |
| 17 | 125610055 | 1,56 | 2,19 | 2,19 | 2,00 | 71 | LT |
| 18 | 125610056 | 1,86 | 1,94 | 2,39 | 1,72 | 73 | LT |
| 19 | 125610069 | 2,33 | 2,67 | 2,35 | 1,26 | 81 | LT |
| 20 | 125610083 | 2,71 | 2,05 | 2,61 | 2,50 | 80 | LT |

**Keterangan Status:**

Lulus Cepat (LC) : 7 semester

Lulus Tepat (LT) : 8 sampai dengan 10 semester

Lulus Lambat (LL) : lebih dari 10 semester

Terdapat data baru mahasiswa sebagai data uji kemudian dilakukan perhitungan berdasarkan metode K –NN dengan *euclidean distance* sebagai berikut :

Nim : 165610122

IPS 1 : 3,83

IPS 2 : 3, 73

IPS 3 : 3,50

IPS 4 : 3,10

Total sks semester 4 : 77

Dari data baru untuk mahasiswa dengan nim 165610122 tersebut dihitung jarak dengan menggunakan rumus (2.2) nilai K adalah 5 sebagai berikut :

1. Nilai d(x21, y1) =

$$\sqrt{(3,83-3,19)^2+(3,73-3,52)^2+(3,50-3,50)^2+(3,10-3,13)^2+(77-91)^2}$$

$= \sqrt{0,4096+0,0441+0+0,0009+196} = \sqrt{196,4546}$

$= 14,0162$

2. Nilai d(x21, y2)

$$= \sqrt{(3,83-3,19)^2+(3,73-3,52)^2+(3,50-3,33)^2+(3,10-3,22)^2+(77-91)^2}$$

$= \sqrt{0,4096+0,0441+0,0289+0,0144+196} = \sqrt{196,4970}$

$= 14,0177$

3. Nilai d(x21, y3)

$= \sqrt{(3,83-3,86)^2+(3,73-3,65)^2+(3,50-3,36)^2+(3,10-3,38)^2+(77-88)^2}$

$= \sqrt{0,0009+0,0064+0,0196+0,3364+121} = \sqrt{121,3633}$

$= 11,0165$

4. Nilai d(x21,y4)

$= \sqrt{(3,83-3,38)^2+(3,73-3,38)^2+(3,50-3,29)^2+(3,10-3,33)^2+(77-93)^2}$

$= \sqrt{0,2025+0,12225+0,0441+0,0529+256} = \sqrt{256,4220}$

$= 16,0132$

Berikut ini adalah tabel hasil perhitungan jarak menggunakan *Euclidean Distance :*

**Tabel 2.3 Perhitungan Jarak menggunakan *Euclidean Distance***

| NO | NIM | JARAK EUCLIDEAN | RANKING |
|---|---|---|---|
| 1 | 135610075 | 14,0162 | 17 |
| 2 | 135610082 | 14,0177 | 18 |
| 3 | 135610123 | 11,0165 | 15 |
| 4 | 135610140 | 16,0132 | 20 |
| 5 | 135610152 | 9,1287 | 13 |
| 6 | 135610157 | 12,0229 | 16 |
| 7 | 085610034 | 8,8692 | 12 |
| 8 | 095610111 | 10,2399 | 14 |
| 9 | 095610116 | 14,8750 | 19 |
| 10 | 105610080 | 7,7235 | 11 |
| 11 | 105610087 | 5,3121 | 5 |
| 12 | 115610022 | 7,6883 | 10 |
| 13 | 115610062 | 6,1637 | 7 |
| 14 | 115610118 | 5,5183 | 6 |
| 15 | 125610042 | 4,7486 | 2 |
| 16 | 125610050 | 6,3851 | 8 |
| 17 | 125610055 | 6,8155 | 9 |
| 18 | 125610056 | 5,1207 | 4 |
| 19 | 125610069 | 4,9073 | 3 |
| 20 | 125610083 | 3,7721 | 1 |

- *Nearest neighbor* ditentukan pada awal adalah k = 5, yaitu 5 jarak yang paling kecil :

**Tabel 2.4 Jarak terdekat sebanyak k = 5**

| No | NIM | Jarak | Rangking | Status |
|---|---|---|---|---|
| 1 | 125610083 | 3,7721 | 1 | LT |
| 2 | 125610042 | 4,7486 | 2 | LT |
| 3 | 125610069 | 4,9073 | 3 | LT |
| 4 | 125610056 | 5,1207 | 4 | LT |
| 5 | 105610087 | 5,3121 | 5 | LL |

- Menghitung jumlah status yang lebih banyak muncul. Pada bagian ini, kemunculan status terbanyak adalah LT sebanyak 4 kali, sedangkan kemunculan status LL sebanyak satu kali.
- Kesimpulan dari hasil perkiraan masa studi mahasiswa dengan NIM 165610122 adalah Lulus Tepat Waktu (LT).