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

Selvia Lorena Br Ginting[1)], Wendi Zarman[2)], Astrid Darmawan[3)]
[1,2,3)] Jurusan Teknik Komputer, Fakultas Teknik dan Ilmu Komputer
Universitas Komputer Indonesia
Jl. Dipatiukur No.112-116 Bandung 40132
e-mail: selvialorena@yahoo.com

ABSTRAK

Data mining adalah sebuah teknik yang memanfaatkan data dalam jumlah yang besar untuk memperoleh informasi berharga yang sebelumnya tidak diketahui dan dapat dimanfaatkan untuk pengambilan keputusan penting. Data mining juga memanfaatkan pengalaman atau bahkan kesalahan di masa lalu untuk meningkatkan kualitas dari model maupun hasil analisanya, salah satunya dengan kemampuan pembelajaran yang dimiliki teknik data mining yaitu klasifikasi. Kegiatan pengklasifikasian yang dilakukan oleh manusia masih memiliki keterbatasan, terutama pada kemampuan manusia dalam menampung jumlah data yang ingin diklasifikasikan. Selain itu bisa juga terjadi kesalahan dalam pengklasifikasian yang dilakukan. Salah satu cara mengatasi masalah ini adalah dengan menggunakan Data Mining (DM) dengan teknik klasifikasi. Klasifikasi merupakan tugas pembelajaran yang memetakan sebuah objek baru ke dalam salah satu label class atau kategori pada objek lama yang telah didefinisikan sebelumnya. Klasifikasi ini menggunakan salah satu metode algoritma data mining yaitu k-Nearest Neighborhood (KNN). Algoritma KNN bekerja berdasarkan jarak terpendek dari objek baru ke objek lama dengan menentukan nilai k. Nilai k merupakan parameter untuk menentukan jarak terdekat antara objek baru terhadap objek lama. Dengan menggunakan teknik data mining tersebut maka di perguruan tinggi dapat memanfaatkan data akademik mahasiswa yaitu indeks prestasi (IP) untuk memprediksi masa studi mahasiswa berdasarkan kategori kelulusan yaitu tepat waktu (4-5 Tahun) dan tidak tepat waktu (5 tahun lebih). Dalam aplikasi data mining ini terdiri dari data testing (data yang akan diuji) dan data training (data yang telah diketahui label class atau kategorinya) dengan masukan NIM dan nilai k. Nilai k yang terbaik, tergantung pada jumlah data yang digunakan. Jika nilai k tinggi, maka hasil tingkat keberhasilannya belum tentu baik dan begitu sebaliknya. Sehingga diharapkan hasil akhir dari aplikasi data mining ini dapat menghasilkan prediksi masa studi mahasiswa.

Kata Kunci: Data Mining, Klasifikasi, Algoritma k-Nearest Neighborhood, Prediksi Masa Studi Mahasiswa

1. PENDAHULUAN

Dalam dunia pendidikan terutama pendidikan tinggi, data yang berlimpah dan berkesinambungan mengenai mahasiswa yang dibina dan alumni terus dihasilkan. Pertumbuhan yang pesat dari pertambahan data akademik ini telah menciptakan kondisi dimana suatu perguruan tinggi memiliki tumpukan data yang banyak. Namun pada saat ini, tumpukan data tersebut banyak yang belum dimanfaatkan secara maksimal bahkan tidak terpakai. Padahal tumpukan data tersebut dapat menjadi sebuah informasi yang bermanfaat dengan menggunakan suatu teknik yaitu teknik *data mining*. *Data mining* adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu *database*. Penggunaan teknik *data mining* pada perguruan tinggi dapat berguna mengolah dan menyebarkan informasi untuk menunjang kegiatan operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis. Data yang berlimpah tersebut membuka peluang diterapkannya *data mining* untuk pengelolaan pendidikan yang lebih baik dan *data mining* dalam pelaksanaan pembelajaran berbantuan komputer yang lebih efektif dalam suatu perguruan tinggi.

Penelitian ini memanfaatkan data akademik yang sebelumnya hanya menjadi beban *database* yang dimiliki oleh jurusan Teknik Komputer UNIKOM, yaitu data IP (Indeks Prestasi) mahasiswa dari semester satu sampai semester enam khususnya Program Sarjana (S1). Data ini akan dimanfaatkan sebagi sumber informasi strategis bagi jurusan untuk memprediksi masa studi mahasiswa dengan menerapkan salah satu teknik dari *data*

mining yaitu klasifikasi dengan algoritma k-*Nearest-Neighborhood*. Hal ini dilakukan dengan harapan dapat menemukan informasi tingkat kelulusan dan persentase kelulusan mahasiswa sehingga dapat digunakan oleh pihak jurusan untuk mencari solusi atau kebijakan dalam proses evaluasi pembelajaran di Jurusan Teknik Komputer. Tujuan yang ingin dicapai dalam pembangunan aplikasi *data mining* ini dapat memprediksi masa studi mahasiswa sehingga dapat mengetahui tingkat kelulusan dan persentase kelulusan mahasiswa di jurusan Teknik Komputer.

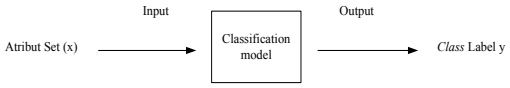
2. LANDASAN TEORI

Data Mining

Data mining merupakan proses menemukan pengetahuan yang menarik dari data yang berjumlah besar yang disimpan di dalam *database*, gudang data atau repositori informasi. *Data mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data. Secara umum, *data mining* dapat melakukan dua hal yaitu memberikan kesempatan untuk menemukan informasi menarik yang tidak terduga dan juga bisa menangani data berskala besar. Dalam menemukan informasi yang menarik ini, ciri khas *data mining* adalah kemampuan pencarian secara hampir otomatis, karena dalam banyak teknik *data mining* ada beberapa parameter yang masih harus ditentukan secara manual atau semi manual. *Data mining* juga dapat memanfaatkan pengalaman atau bahkan kesalahan di masa lalu untuk meningkatkan kualitas dari model maupun hasil analisanya, salah satunya dengan kemampuan pembelajaran yang dimiliki beberapa teknik *data mining* seperti klasifikasi.

Klasifikasi

Klasifikasi adalah tugas pembelajaran sebuah fungsi target f yang memetakan setiap himpunan atribut x ke salah satu label *class* y yang telah didefinisikan sebelumnya. Klasifikasi dapat juga diartikan suatu proses untuk menemukan suatu model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan dapat menggunakan model untuk memprediksi kelas objek yang label *class*-nya tidak diketahui.



Gambar 1. Model Klasifikasi

Data input untuk klasifikasi adalah isi dari *record*. Setiap *record* dikenal sebagai *instance* atau contoh, yang ditentukan oleh sebuah *tuple* (x, y), dimana x adalah himpunan atribut dan y adalah atribut tertentu, yang dinyatakan sebagai label *class* (juga dikenal sebagai kategori atau atribut target).

Pendekatan umum yang digunakan dalam masalah klasifikasi adalah pertama, *data testing* berisi *record* yang mempunyai label *class* yang telah diketahui. *Data training* digunakan untuk membangun model klasifikasi yang kemudian diaplikasikan ke *data testing* yang berisi *record-record* dengan label *class* yang tidak diketahui.

Algoritma Nearest Neighborhood

Algoritma *Nearest Neighborhood* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru (*data testing*) dengan kasus lama (*data training*), yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Jenis algoritma *Nearest Neighborhood* ada 2, yaitu:

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

Di dalam penelitian ini akan digunakan Algoritma *k-Nearest Neighborhood.*

Algoritma k-Nearest Neighborhood (k-NN)

Algoritma *k-Nearest Neighborhood* (*k*-NN) adalah suatu metode yang menggunakan algoritma *supervised* dimana hasil dari *query instance* yang baru diklasifikasi berdasarkan mayoritas dari label *class* pada *k*-NN. Tujuan dari algoritma *k*-NN adalah mengklasifikasi objek baru berdasarkan atribut dan *data training*.

Algoritma *k*-NN bekerja berdasarkan jarak terpendek dari *query instance* ke *data training* untuk menentukan *k*-NN-nya. Salah satu cara untuk menghitung jarak dekat atau jauhnya tetangga menggunakan metode *euclidian distance.*

Ecludian Distance sering digunakan untuk menghitung jarak. *Euclidian Distance* berfungsi menguji ukuran yang bisa digunakan sebagai interpretasi kedekatan jarak antara dua obyek, di bawah ini merupakan rumus *Ecludian Distance*:

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

Dimana,

X_{ik} = nilai X pada *data training*
X_{jk} = nilai X pada *data testing*
m = batas jumlah banyaknya data

Jika hasil nilai dari rumus di atas besar maka akan semakin jauh tingkat keserupaan antara kedua objek dan sebaliknya jika hasil nilainya semakin kecil maka akan semakin dekat tingkat keserupaan antar objek tersebut. Objek yang dimaksud adalah *data training* dan *data testing*.

Dalam algoritma ini, nilai *k* yang terbaik itu tergantung pada jumlah data. Ukuran nilai *k* yang besar belum tentu menjadi nilai *k* yang terbaik begitupun juga sebaliknya.

Langkah-langkah untuk menghitung algoritma *k*-NN:

1. Menentukan nilai *k*.
2. Menghitung kuadrat jarak *euclid* (*query instance*) masing-masing objek terhadap *data training* yang diberikan.
3. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak *euclid* terkecil.
4. Mengumpulkan label *class* Y (klasifikasi *Nearest Neighborhood*).
5. Dengan menggunakan kategori *Nearest Neighborhood* yang paling mayoritas maka dapat dipredeksikan nilai *query instance* yang telah dihitung.

3. ANALISIS DAN PERANCANGAN PERANGKAT LUNAK

Analisis Data

UNIKOM merupakan salah satu perguruan tinggi swasta di Bandung. UNIKOM memiliki beberapa Jurusan salah satunya adalah Jurusan Teknik Komputer. Jurusan ini termasuk kategori yang sangat sulit untuk lulus tepat waktu. Setiap tahun, Jurusan Teknik Komputer hanya menghasilkan beberapa mahasiswa yang lulus 4 atau 5 tahun. Karena jumlah kelulusan tiap tahunnya hanya sedikit, maka peneliti memanfaatkan data nilai IP mahasiswa di Jurusan Teknik Komputer untuk menemukan informasi atau pengetahuan baru yang berguna dalam mengambil sebuah keputusan dan membantu dalam evaluasi sistem pembelajaran di Jurusan Teknik Komputer. Informasi yang dibutuhkan adalah memprediksi masa studi mahasiswa dengan atribut IP dari semester satu sampai semester enam.

Data akademik mahasiswa yang diambil adalah data mahasiswa angkatan 2001-2006. Hal ini didasarkan pada kebutuhan data yang akan dihubungkan dengan *data testing*, dengan asumsi bahwa mahasiswa angkatan 2001-2006 akan lulus dari rentang waktu tahun 2005-2010. Sedangkan data kelulusan dalam *data training* rentang waktunya dari tahun 2004-2011.

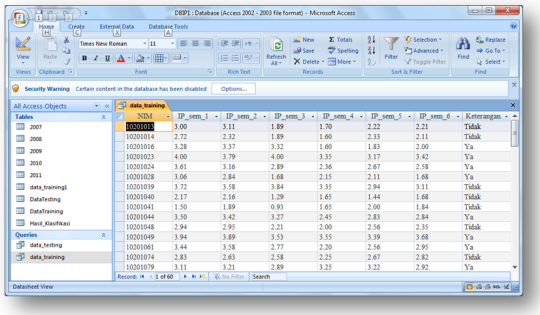
Aplikasi data mining yang dibuat terdiri dari dua data, yaitu:

1. *Data Testing*
   a. NIM
   b. Indeks Prestasi (IP) mahasiswa dari semester satu sampai enam.
2. *Data Training*
   a. NIM
   b. Indeks Prestasi (IP) mahasiswa dari semester satu sampai enam.
   c. Keterangan (Kategori Kelulusan)

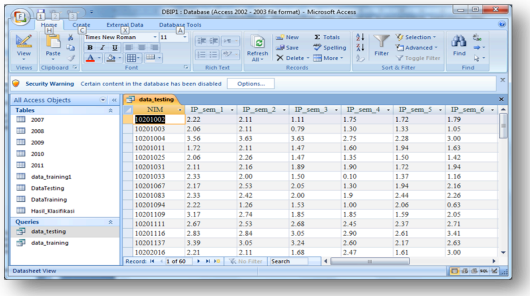
Data training memiliki kategori sebagai berikut:

Tabel 1. Kategori Kelulusan Berdasarkan Lama Studi

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

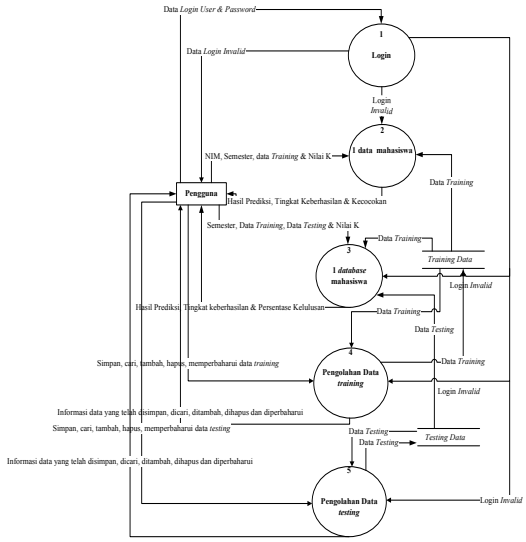


Gambar 2. Cuplikan *Data Training* (61 data)

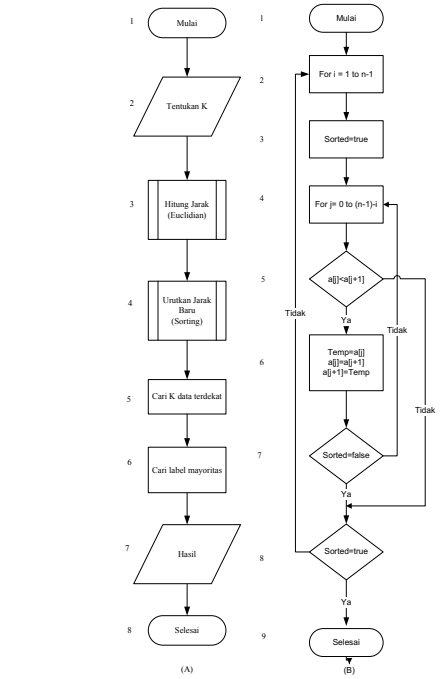


Gambar 3. Cuplikan *Data Testing* (30 data)

Data Flow Diagram Sistem



Flowchart Sistem



Gambar 4. (A) Algoritma KNN & (B) Algoritma Sorting

Tabel 2. Penjelasan Flowchart Algoritma *k*-NN

Urutan	Keterangan
1	Memulai awal program.
2	Menentukan nilai *k*.
3	Pemanggilan procedure untuk menghitung jarak baru menggunakan rumus *euclidean*.
4	Pemanggilan procedure untuk mengurutkan jarak baru menggunakan algoritma *sorting*, yaitu *insertion*.
5	Mencari jarak terdekat sesuai nilai *k*.
6	Mencari mayoritas label *class* pada jarak terdekat sesuai nilai *k*.
7	Menghasilkan prediksi.
8	Program selesai.

Tabel 3. Penjelasan Flowchart Algoritma *Sorting*

Urutan	Keterangan
1	Memulai *sorting*.
2	Pengulangan inisialisasi i dari 1 hingga (n-1).
3	*Sorting* inisialisasi i benar.
4	Pengulangan inisialisasi j dari 0 hingga ((n-1)-i).
5	Apakah isi a[j]<a[j+1] ? Jika Ya, maka melakukan pertukaran data dan jika Tidak, maka melanjutkan proses di urutan 8.
6	Melakukan pertukaran isi data yang terkecil hingga terbesar.
7	Jika *sorting*-nya salah, maka kembali ke urutan 4.
8	Apakah *sorting*-nya sudah benar? Jika Tidak, maka mengulang proses urutan 2 dan jika Ya, maka *sorting* selesai
9	Proses *sorting* selesai.

4. PENGUJIAN SISTEM

Pengujian berguna untuk mengukur kehandalan dari sistem atau alat yang dibangun, sehingga hasil yang diharapkan dapat sesuai dengan yang dibutuhkan.

Pengujian dilakukan terhadap aplikasi *data mining* yang dibangun untuk melihat apakah aplikasi ini berhasil atau tidak dalam memprediksi masa studi mahasiswa. Pengujian terdiri dari 2 proses yaitu:

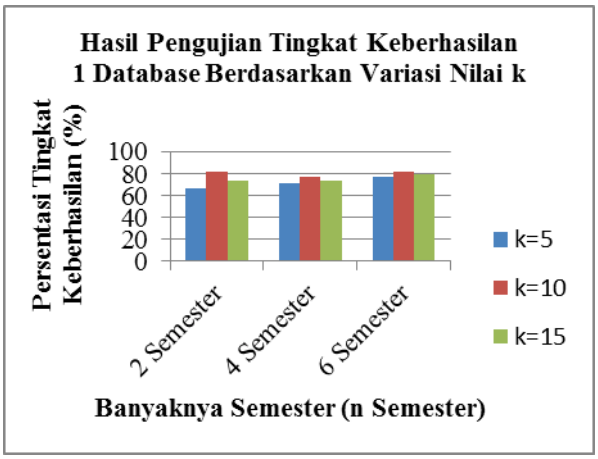
1. Pengujian 1 *database* mahasiswa (keseluruhan) dengan menggunakan *data training* yang berjumlah 30 data.
2. Pengujian 1 *database* mahasiswa (keseluruhan) dengan menggunakan *data training* yang berjumlah 61 data.

Masing-masing proses pengujian tersebut menggunakan Indeks Prestasi (IP) dua semester (semester 1 dan 2), empat semester (semester 1-4) dan enam semester (semester 1-6) dengan menggunakan nilai k yang berbeda. Pengujian ini dilakukan dengan menggunakan *data training* yang berjumlah sebanyak 61 data dan 30 data serta *data testing* yang berjumlah 60 data. Hasil prediksi 1 *database* akan dibandingkan dengan data asli dan dicari kecocokannya secara otomatis oleh program.

Pengujian ini juga berguna untuk mengetahui apakah nilai k yang digunakan adalah nilai k yang terbaik dengan hasil tingkat keberhasilannya tinggi atau tidak untuk memprediksi kelulusan mahasiswa pada sistem aplikasi *data mining* ini.

Untuk mengetahui tingkat keberhasilan pada sistem ini digunakan rumus:

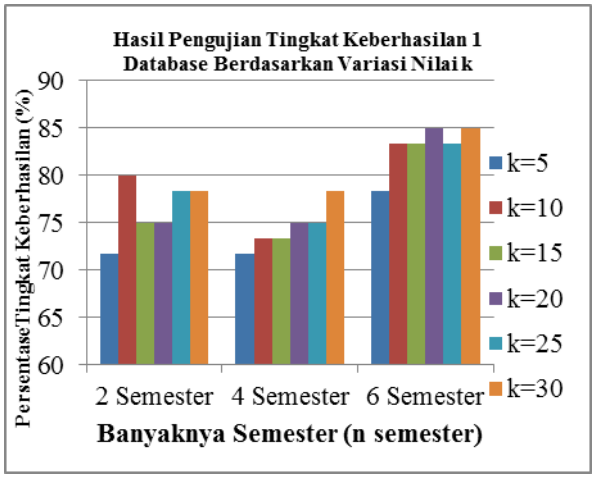
$$\frac{\sum \textit{hasil pengujian bernilai benar}}{\sum \textit{banyaknya data sampel}} \; x\ 100\%$$



Gambar 5. Grafik Pengujian 1 *database* dengan *Data Training* = 30 data

Dari hasil gambar 5 dapat disimpulkan sebagai berikut:

1. Untuk dua semester yaitu nilai k yang terbaik untuk digunakan memprediksi studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 81.66%.
2. Untuk empat semester yaitu nilai k yang terbaik untuk digunakan memprediksi studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 76.66%.
3. Untuk enam semester yaitu nilai k=10 merupakan nilai k yang terbaik untuk digunakan memprediksi masa studi mahasiswa dengan tingkat keberhasilan 81.66%.
4. Masing-masing nilai k terbaik di atas hanya berlaku untuk *data training* yang berjumlah 30data dan *data testing*-nya 60 data.



Gambar 6. Grafik Pengujian 1 *database* dengan *Data Training* = 61 data

Dari hasil Gambar 6 dapat disimpulkan sebagai berikut:

1. Untuk dua semester yaitu nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 80%.
2. Untuk empat semester, nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=30 dengan tingkat keberhasilan 78.33%.
3. Untuk enam semester, nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=20 dan k=30 dengan tingkat keberhasilan 85%.
4. Masing-masing nilai k terbaik di atas hanya berlaku untuk *data training* yang berjumlah 61 data dan *data testing*-nya 60 data.

Pada percobaan satu dan dua di atas dapat dilihat bahwa dengan mengubah nilai k akan menghasilkan data yang bervariasi. Jadi, ukuran nilai k yang besar untuk digunakan memprediksi masa studi mahasiswa belum tentu menjadi nilai k yang terbaik dengan tingkat keberhasilan yang tinggi begitupun

juga sebaliknya. Nilai k yang terbaik dipengaruhi oleh jumlah data yang digunakan.

5. KESIMPULAN

Berdasarkan pemaparan diatas, dapat diambil beberapa kesimpulan, yaitu:

1. *Data training* dengan jumlah 30 data digunakan untuk menguji *data testing* berjumlah 30 data, didapatkan nilai k yang terbaik untuk memprediksi masa studi mahasiswa yaitu sebagai berikut:
   1) Untuk dua semester yaitu nilai k yang terbaik untuk digunakan memprediksi studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 81.66%.
   2) 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=10 merupakan nilai k yang terbaik untuk digunakan memprediksi masa studi mahasiswa dengan tingkat keberhasilan 81.66%.
2. *Data training* dengan jumlah 60 data digunakan untuk menguji *data testing* berjumlah 61 data, diperoleh nilai k yang terbaik untuk memprediksi masa studi mahasiswa yaitu sebagai berikut:
   1) Untuk dua semester yaitu nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=10 dengan tingkat keberhasilan 80%.
   2) Untuk empat semester, nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=30 dengan tingkat keberhasilan 78.33%.
   3) Untuk enam semester, nilai k yang terbaik untuk memprediksi masa studi mahasiswa adalah nilai k=20 dan k=30 dengan tingkat keberhasilan 85%.

6. DAFTAR PUSTAKA

[1] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques*. Second Edition. San Fransisco: Morgan Kauffman.

[2] Huda, Masykur, N. (2010). *Aplikasi Data Mining untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa.* Semarang: Universitas Diponegoro.

[3] Keller, J.M., Gray, M.R., & Givens, J.A (1985). *A Fuzzy k-Nearest Neighbor Algorithm.* Systems, Man and Cybermatics, IEEE Transactions. (4). 580-585.

[4] Kusrini & Luthfi, E., T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi.

[5] Rismawan, Tedy, et al. (2008). *Sistem Pendukung Keputusan Berbasis Pocket PC Sebagai Penentu Status Gizi Menggunakan Metode KNN (K-Nearest Neighbor)*. Jurnal Teknoin.

[6] Seidl, Thomas, and Kriegel, H. P. (1998). *Optimal Multi-Step k-Nearest Neighbor Search.* ACM SIGMOD Record. Vol 27. No. 2. ACM.