

BAB II

TINJAUAN PUSTAKA

Bab ini berisi pembahasan mengenai kajian pustaka dan teori dasar yang mendukung dan berhubungan dengan penelitian. Kajian pustaka membahas penelitian yang telah ada dan diusulkan, sedangkan dasar teori berisi teori-teori apa saja yang berguna dan diperlukan dalam penyusunan penelitian. Kajian pustaka yang dipakai sebagai acuan penelitian ini berjudul *Web Document Classification Based on Fuzzy K-NN Algorithm* milik Juan Zang, dkk. Dasar teori yang akan digunakan meliputi klasifikasi, *data mining*, logika *fuzzy*, *k-nearest neighbor* dan *fuzzy k-nearest neighbor*.

2.1 Kajian Pustaka

Dalam jurnal yang diterbitkan oleh International Conference on Computational Intelligence and Security pada tahun 2009 tersebut menyimpulkan bahwa hasil eksperimen menggambarkan bahwa performa klasifikasi dari FK-NN lebih baik daripada K-NN ataupun *Support Vector Machine* (SVM), tetapi kecepatan klasifikasinya lebih lambat daripada K-NN. Hasil akurasi dari perbandingan ketiga metode tersebut tersaji dalam Tabel 2.1 [ZAN-09].

Tabel 2.1 Perbandingan performa klasifikasi dalam *fuzzy* K-NN, K-NN dan SVM

Algoritma	Ukuran Vektor				
	500	1000	1500	2000	2500
Fuzzy K-NN	64.12	70.23	72.17	74.98	76.38
K-NN	58.23	63.34	64.2	65	65.13
SVM	58.45	60.7	61.35	64.42	67.27

Sumber : [ZAN-09]

Tabel 2.1 menunjukkan skala perbandingan performa klasifikasi antar metode. Ukuran tersebut diperoleh dengan perhitungan menggunakan Persamaan 2.3. Tabel 2.1 memperlihatkan bahwa performa metode F-KNN lebih baik dibanding metode lain dalam hal klasifikasi dokumen web dengan melibatkan ukuran vektor berapapun.

Perbedaan penelitian ini dengan penelitian yang menjadi acuan adalah dalam hal kasus yang diteliti. Jika dokumen acuan lebih kepada klasifikasi dokumen web, penelitian ini mengacu pada klasifikasi kategori bayar dengan lingkup Universitas Brawijaya Malang dengan data numerik.

2.2 Gambaran Umum Seleksi Masuk Mahasiswa Baru

Seleksi masuk mahasiswa baru Universitas Brawijaya Malang tahun ajaran 2013/2014 mempunyai beberapa jalur. Dikutip dari situs resmi seleksi masuk mahasiswa baru www.selma.ub.ac.id terdapat beberapa jalur masuk, yaitu jalur Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN), jalur Seleksi Bersama Nasional Masuk Perguruan Tinggi Negeri (SBNMPTN), jalur mandiri Seleksi Program Minat dan Kemampuan (SPMK), jalur Seleksi Program Khusus Penyandang Disabilitas (SPKPD), jalur masuk *International Student Admission* dan jalur seleksi alih program [SEL-13]. Dalam penelitian ini pembahasan hanya sebatas pada seleksi jalur masuk SNMPTN, SPMK dan SBNMPTN.

SNMPTN 2013 merupakan pola seleksi nasional yang dilaksanakan oleh Panitia Pelaksana SNMPTN 2013 yang ditetapkan oleh Direktur Jenderal Pendidikan Tinggi Kementerian Pendidikan dan Kebudayaan dan diikuti oleh seluruh Perguruan Tinggi Negeri dalam satu sistem yang terpadu. Biaya pelaksanaan SNMPTN 2013 ditanggung oleh Pemerintah, sehingga peserta tidak dipungut biaya pendaftaran. Sejalan dengan program Pemerintah tentang Bidikmisi, peserta dari keluarga kurang mampu secara ekonomi dan mempunyai prestasi akademik tinggi dapat mengikuti seleksi melalui SNMPTN dan jika diterima PTN akan mendapatkan beasiswa selama masa studi normal [SEL-13].

SBNMPTN (Seleksi Bersama Masuk Perguruan Tinggi Negeri) merupakan pola seleksi nasional berdasarkan hasil tes tertulis / ujian keterampilan, dengan biaya pendaftaran ditanggung oleh peserta. SBNMPTN dilaksanakan secara serentak di 62 PTN [SEL-13].

SPMK (Seleksi Program Minat dan Kemampuan) merupakan seleksi penerimaan mahasiswa UB melalui ujian tulis yang dilakukan khusus oleh Universitas Brawijaya [SEL-13].

Di luar penjelasan tentang beberapa jalur masuk mahasiswa baru Universitas Brawijaya tersebut di atas, penelitian ini lebih mengarah pada besarnya biaya masuk melalui jalur SNMPTN, SPMK dan SBNMPTN. Rektor Universitas Bawijaya dalam pengumumannya tanggal 31 Juli 2013 menyampaikan bahwa Pembayaran Uang Kuliah (UKT) bagi Calon Mahasiswa Baru Jalur SNMPTN, SBNMPTN dan SPMK dibayarkan satu tahun penuh yaitu untuk semester Ganjil dan Genap Tahun Akademik 2013/2014 serta tidak ada penundaan, keringanan atau angsuran [SEL-13].

Dalam hal UKT Rektor Universitas Brawijaya telah menetapkan sebanyak enam kategori pembayaran bagi jalur masuk SBNMPTN dan SNMPTN. Untuk Program Teknologi Informasi dan Ilmu Komputer itu sendiri kategori satu memiliki nominal Rp 500.000,00, kategori dua sebesar Rp 1.000.000,00, kategori tiga sebesar Rp 6.500.000,00, kategori empat sebesar Rp 7.500.000,00, kategori lima sebesar Rp 8.500.000,00 dan kategori enam sebesar Rp 9.500.000,00 [SEL-13]. Rincian biaya di atas dibayarkan sekaligus untuk tahun pertama, tidak ada pungutan pembayaran lain sampai lulus (max 8 semester) dan pada tahun berikutnya dibayar tiap semester.

Dalam hal jalur masuk SPMK, Rektor Universitas brawijaya menetapkan sebanyak tiga jenis kategori SPP yang nominalnya berbeda antara program studi teknik informatika, sistem komputer dan sistem informasi, yaitu kategori satu teknik informatika memiliki nominal Rp 4.500.000, kategori dua Rp 5.750.000 dan kategori tiga Rp 6.750.000. Sedangkan program studi sistem komputer dan sistem informasi kategori satunya memiliki nominal Rp 4.000.000, kategori dua Rp 5.000.000 dan kategori tiga Rp 6.000.000. Biaya-biaya tersebut belum termasuk DBP, *bandwidth* dan SPFP [SEL-13].

2.3 Klasifikasi

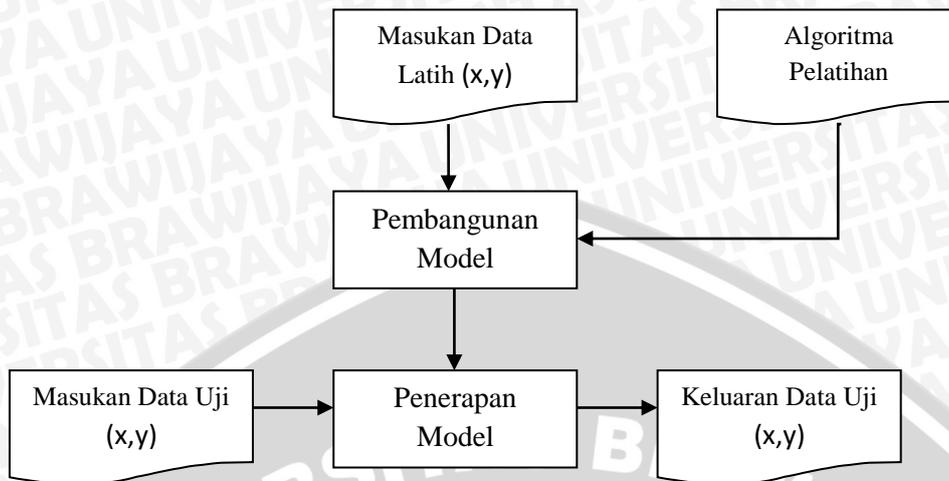
Menurut Kamus Besar Bahasa Indonesia (KBBI), klasifikasi merupakan kata benda yang berarti penyusunan bersistem di kelompok atau golongan menurut kaidah atau standar yang ditetapkan [KBB-12]. Arti yang lebih spesifik dari klasifikasi yang berkaitan dengan kasus yang dikaji dalam penelitian ini seperti diungkap oleh Feri Sulistianta dan Dominikus Juju dalam bukunya yang

berjudul *Data Mining*, bahwa klasifikasi adalah teknik yang digunakan untuk memprediksi keanggotaan kelompok untuk contoh data [SUD-10].

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu (1) pembangunan model sebagai *prototype* untuk disimpan sebagai memori dan (2) penggunaan model tersebut untuk melakukan pengenalan / klasifikasi / prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya. Selain itu klasifikasi juga dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan / pembelajaran terhadap fungsi target f yang memetakan setiap set atribut (fitur) x ke satu dari sejumlah label kelas y yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori [PRA-12].

Model dalam klasifikasi mempunyai arti yang sama dengan kotak hitam, di mana ada suatu model yang menerima masukan, kemudian mampu melakukan pemikiran terhadap masukan tersebut, dan memberikan jawaban sebagai keluaran dari hasil pemikirannya. Kerangka kerja klasifikasi ditunjukkan pada Gambar 2.1. Pada gambar tersebut disediakan sejumlah data latih (x,y) untuk digunakan sebagai data pembangun model. Model tersebut kemudian dipakai untuk memprediksi kelas dari data uji $(x,?)$ sehingga diketahui kelas y yang sesungguhnya [PRA-12].

Dari kedua pengertian di atas terdapat kemiripan makna antara satu dengan yang lain, yaitu bahwa dalam melakukan klasifikasi dibutuhkan sebuah metode khusus yang berguna sebagai pembatas kelas, sehingga dalam setiap kelas memiliki anggota yang mempunyai kemiripan. Sedangkan antar kelas memiliki ciri atau keunikan tersendiri. Sebuah metode berguna untuk mencari kemiripan suatu individu atau data terhadap kelas-kelas yang tersedia.



Gambar 2.1 Proses Pekerjaan Klasifikasi

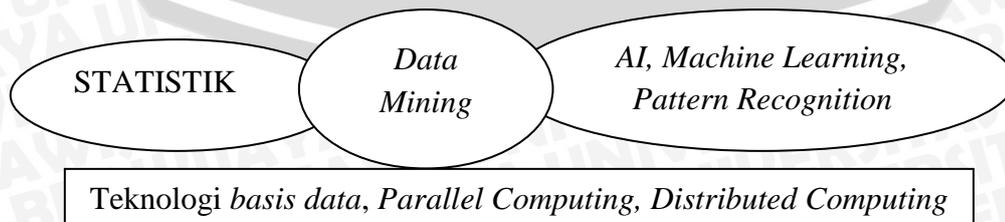
Sumber : [PRA-12]

Klasifikasi dalam penelitian ini nantinya akan memilah dan memilih kategori bayar mahasiswa baru menggunakan metode *Fuzzy K-Nearest Neighbor*. Untuk kategori itu sendiri pihak Universitas Brawijaya memiliki sebanyak enam kategori bayar yang masing-masing memiliki syarat tertentu dalam penempatannya. Diharapkan dengan adanya penelitian ini dapat membantu pihak orang tua wali mahasiswa Universitas Brawijaya khususnya Program Teknologi Informasi dan Ilmu Komputer dalam dalam hal keuangan.

2.4 Data Mining

2.4.1 Pengertian Data Mining

Data mining merupakan proses pencarian pola yang menarik dan pengetahuan dari sejumlah data yang banyak. Sumber data itu sendiri dapat berupa basis data, gudang data, web, gudang informasi lain, atau data yang dialirkan ke dalam sistem dinamis [HMJ-12].



Gambar 2.2 Posisi Data Mining Di antara Beberapa Bidang Ilmu

Sumber : [PRA-12]

Data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah *data mining* kadang disebut juga *knowledge discovery*. Salah satu teknik dalam *data mining* adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali prediksi juga dapat memanfaatkan teknik ini [PRA-12].

2.4.2 Pekerjaan dalam Data Mining

Pekerjaan dalam *data mining* dapat dibagi menjadi empat kelompok, yaitu model prediksi (*prediction modelling*), analisis kelompok (*cluster analysis*), analisis asosiasi (*association analysis*), dan deteksi anomali (*anomaly detection*) [PRA-12]. Oleh karena penelitian masuk dalam analisis kelompok, maka pembahasan akan terfokus hanya pada analisis kelompok saja.

Analisis kelompok melakukan pengelompokan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data-data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, dan akan terpisah dalam kelompok yang berbeda jika keluar dari batas kesamaan dengan kelompok tersebut [PRA-12].

Penelitian ini termasuk dalam kategori analisis kelompok karena penelitian ini memiliki kerangka kerja yang sama dengan konsep analisis kelompok, yaitu melakukan pengelompokan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan kesamaan karakteristik data. Dalam hal ini, penelitian merujuk pada pengklasifikasian kategori bayar mahasiswa baru [PRA-12].

2.4.3 Metode Data Mining

Data mining terdiri dari beberapa metode. Metode yang berbeda menyajikan tujuan yang berbeda, masing-masing metode menawarkan kelebihan dan kekurangan tersendiri. Dalam *data mining*, klasifikasi adalah salah satu tugas yang sangat penting. Itu memetakan data ke dalam target yang sudah dikenal. Metode yang biasa digunakan untuk tugas klasifikasi *data mining* dapat

dikelompokkan ke dalam beberapa kelompok, yaitu pohon keputusan (*decision trees*), mesin vektor pendukung (*support vektor machines*), algoritma genetika, himpunan *fuzzy*, jaringan saraf (*neural networks*) dan himpunan kasar (*rough set*) [GUP-11]. Oleh karena metode dalam penelitian ini memakai himpunan *fuzzy* dan masuk dalam klasifikasi maka pembahasan akan terfokus pada poin himpunan *fuzzy* dan himpunan kasar (*rough set*) saja.

Bentuk himpunan *fuzzy* sebuah kunci metodologi untuk mewakili dan memproses keraguan. Keraguan timbul dalam banyak bentuk di basis data sekarang ini : ketidaktepatan, ketidaktegasan, ketidakkonsistenan, kesamaran, dan lain-lain. Himpunan *fuzzy* mengeksploitasi keraguan sebagai usaha untuk membuat kerumitan sistem dikendalikan. Selain itu, himpunan *fuzzy* merupakan pendekatan yang tepat untuk menguraikan tak hanya dengan ketidaklengkapan, data *noise* atau tidak tepat, tetapi mungkin juga berguna dalam mengembangkan model yang tidak pasti dari data yang menyediakan performa lebih cerdas dan lebih halus dari pada sistem tradisional [GUP-11].

Sebuah himpunan kasar dapat ditentukan dengan batas atas dan bawah dari himpunan. Setiap anggota dari batas bawah adalah anggota khusus dari himpunan. Setiap bukan anggota dari batas atas adalah bukan anggota khusus dari himpunan. Batas atas dari himpunan kasar adalah perpaduan antara batas bawah dan juga disebut wilayah batas. Anggota dari wilayah batas mungkin (tetapi tidak yakin) anggota dari himpunan. Meskipun, himpunan kasar mungkin dipandang dengan fungsi keanggotaan tiga nilai (iya, tidak, mungkin). Himpunan kasar adalah konsep matematika yang berhadapan dengan data tidak pasti. Mereka biasanya dikombinasi dengan metode lain seperti kaidah induksi, klasifikasi, atau metode klastering [GUP-11].

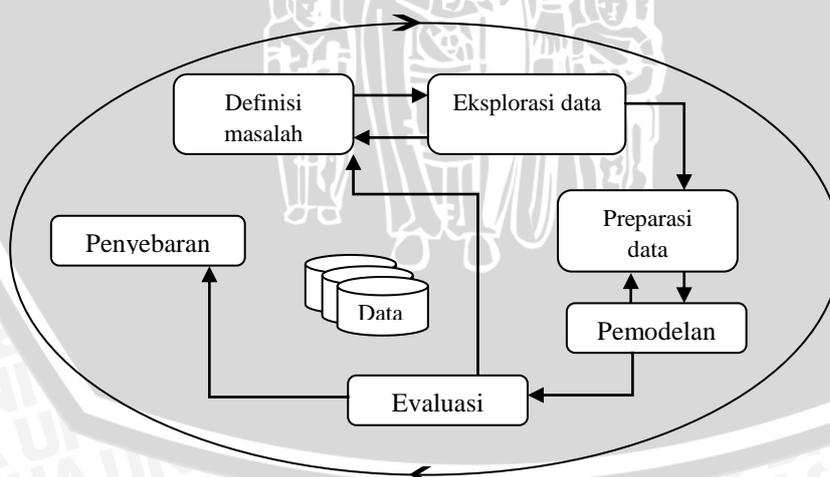
2.4.4 Proses Data Mining

Dalam proses *Knowledge Discovery Databases* (KDD), metode *data mining* adalah untuk ekstraksi pola dari data. Pola tersebut dapat ditemukan tergantung dari tugas *data mining* yang diterapkan. Biasanya, ada dua tipe dari tugas *data mining* : tugas *data mining* deskriptif yang menjelaskan kekayaan umum dari data yang sudah ada, dan tugas *data mining* deskriptif yang mencoba untuk melakukan

prediksi berdasar data yang tersedia. *Data mining* dapat diselesaikan kepada data yang mana kuantitatif, tekstual, atau bentuk multimedia [GUP-11].

Aplikasi *data mining* dapat digunakan parameter jenis berbeda untuk memeriksa data. Mereka meliputi asosiasi (pola yang mana satu peristiwa terkoneksi ke peristiwa lain), klasifikasi (identifikasi dari pola baru dengan target khusus) dan klastering (pengelompokan objek yang identik atau sama). *Data mining* melibatkan diri dalam beberapa langkah di bawah ini [GUP-11] :

- 1) Definisi masalah : Langkah pertama untuk identifikasi tujuan. Berdasarkan penetapan tujuan, rangkaian tepat dari perangkat dapat diaplikasikan pada data untuk membangun model kemiripan perilaku.
- 2) Eksplorasi data : Jika kualitas data tidak cocok untuk model akurasi kemudian rekomendasi koleksi dari data masa depan dan strategi penyimpanan dapat dibuat padanya. Untuk analisa, semua data butuh padu jadi itu dapat diperlakukan secara konsisten.
- 3) Preparasi data : Manfaat dari langkah ini adalah untuk membersihkan dan transformasi data jadi nilai yang hilang dan tidak valid diperlakukan dan semua nilai valid yang diketahui membuat konsisten untuk analisa lebih lanjut.



Gambar 2.3 Representasi Proses *Data Mining*

Sumber : [GUP-11]

- 4) Pemodelan : Berdasarkan data dan keluaran yang diinginkan, algoritma *data mining* dipilih untuk analisa. Algoritma tersebut meliputi teknik klasik seperti statistik, lingkungan dan klastering tetapi juga teknik generasi selanjutnya

seperti pohon keputusan, jaringan dan algoritma berbasis aturan. Algoritma yang spesifik dipilih berdasar tujuan khusus untuk dicapai dan kualitas dari data dianalisa.

- 5) Evaluasi dan penyebaran : Berdasar hasil dari algoritma *data mining*, analisa diadakan untuk menentukan kesimpulan kunci dari analisa dan membuat rangkaian rekomendasi untuk pertimbangan.

2.5 Logika Fuzzy

Logika *fuzzy* adalah aturan atau sistem dari logika. Seperti logika Boolean, logika *fuzzy* berdasar atas teori himpunan. Himpunan dalam logika *fuzzy*, bagaimanapun, memiliki derajat keanggotaan. perubahan pelalaian dalam definisi dari himpunan mempunyai implikasi yang signifikan dan jauh mencapai untuk representasi pengetahuan dan rangkaian dari sebab. Dari perspektif model, logika *fuzzy* adalah logika dari variabel kontinu [HMJ-12].

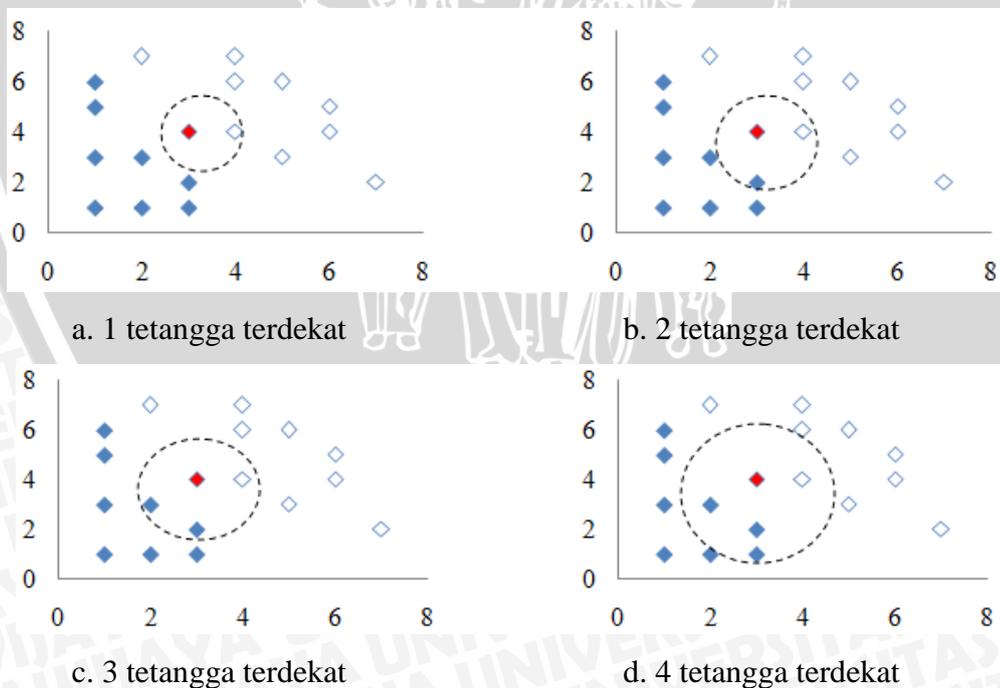
Konsep logika *fuzzy* dicetuskan oleh Lotfi Zadeh, seorang profesor University of California di Berkeley, dan dipresentasikan bukan sebagai metodologi kontrol, namun sebagai suatu cara pemrosesan data yang memperbolehkan anggota himpunan parsial daripada anggota himpunan kosong atau non-anggota. Pendekatan ini pada teori himpunan tidak diaplikasikan untuk mengontrol sistem sampai tahun 70-an karena kurangnya kemampuan komputer mini pada saat itu. Profesor Zadeh beralasan bahwa masyarakat tidak butuh ketepatan, masukan informasi numerik, dan mereka belum sanggup dengan kontrol adaptif yang tinggi. Jika kembalian dari kontroler dapat diprogram untuk menerima *noisy*, masukan yang tidak teliti, mereka akan lebih efektif dan lebih mudah diimplementasikan [ZIM-01].

2.6 K-Nearest Neighbor

2.6.1 Definisi

Algoritma *Nearest Neighbor* (kadang disebut *K-Nearest Neighbor* / K-NN) merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. Prinsip sederhana yang diadopsi oleh algoritma K-NN adalah “Jika suatu hewan berjalan seperti bebek, bersuara kwek-kwek, dan penampilannya seperti bebek, hewan itu mungkin bebek” [PRA-12].

Pada algoritma K-NN, data berdimensi q , jarak dari data tersebut ke data yang lain dapat dihitung. Nilai jarak inilah yang digunakan sebagai nilai kedekatan / kemiripan antara data uji dengan data latih. Nilai K pada K-NN berarti K-data terdekat dari data uji. Gambar 2.5 memberikan contoh algoritma K-NN, bentuk berwarna biru untuk kelas 0, bentuk berwarna putih untuk kelas 1. Seperti yang ditunjukkan ada Gambar 2.5 (a), jika K bernilai 1, kelas dari satu data latih sebagai tetangga terdekat (terdekat pertama) dari data uji tersebut akan diberikan kelas untuk data uji, yaitu kelas 1; jika K bernilai 2, akan diambil 2 tetangga terdekat dari data latih. Begitu juga jika nilai K adalah 3,4,5, dan sebagainya. Jika dalam K-tetangga ada dua kelas yang berbeda, akan diambil kelas dengan jumlah data terbanyak (voting mayoritas), seperti yang ditunjukkan pada Gambar 2.5 (c) dan (d). Pada gambar tersebut terlihat bahwa kelas 0 mempunyai jumlah lebih banyak daripada kelas 1 sehingga data uji akan dikategorikan ke dalam kelas 0. Jika kelas dengan data terbanyak ada dua atau lebih, akan diambil kelas dari data dengan jumlah yang sama tersebut secara acak, seperti yang ditunjukkan pada Gambar 2.5 (c) [PRA-12].



Gambar 2.4 K-NN dengan nilai K-tetangga : (a) 1-NN, (b) 2-NN, (c) 3-NN, (d) 4-NN

Sumber : [PRA-12]

2.6.2 Algoritma *Nearest Neighbor*

Algoritma prediksi K-NN biasanya dilakukan dengan beberapa langkah pasti, seperti :

- 1) $z = (x', y')$, adalah data uji dengan vektor x' dan label kelas y' yang belum diketahui
- 2) Hitung jarak (x', y') , jarak di antara data uji z ke setiap vektor data latih, simpan dalam D
- 3) Pilih $D_z \subseteq D$, yaitu K tetangga terdekat dari z

Pada algoritma tersebut ada sebuah data uji $z = (x', y')$, di mana x' adalah vektor / atribut data uji, sedangkan y' adalah label kelas data uji yang belum diketahui. Hitung jarak (atau kemiripan) data uji ke setiap data latih $d(x', x)$, kemudian ambil K -tetangga terdekat pertama dalam D_z . Setelah itu, hitung jumlah data yang mengikuti kelas yang ada dari K -tetangga tersebut. Kelas dengan data terbanyak yang mengikutinya menjadi kelas pemenang yang diberikan sebagai label kelas pada data uji y' [PRA-12].

Salah satu masalah yang dihadapi K-NN adalah pemilihan K yang tepat. cara voting mayoritas dari K -tetangga untuk nilai K yang besar bisa mengakibatkan distorsi data yang besar. Untuk menangani voting mayoritas tersebut biasanya ditambahkan penggunaan bobot untuk menghitung kandidat kelas yang sebaiknya diambil oleh data uji dari K -tetangga terdekat. Bobot dari setiap tetangga terdekat dihitung dengan [PRA-12] :

$$W_i = \frac{1}{d(x', x_i)^2} \quad (2.1)$$

2.6.3 Karakteristik Klasifikasi dengan *K-Nearest Neighbor*

Nearest neighbor merupakan teknik klasifikasi yang sederhana, tetapi mempunyai hasil kerja yang cukup bagus. Meskipun begitu, K-NN juga mempunyai kelebihan dan kekurangan. Beberapa karakteristik K-NN adalah sebagai berikut [PRA-12] :

- 1) K-NN merupakan algoritma yang menggunakan seluruh data latih untuk melakukan proses klasifikasi (*complete storage*). Hal ini mengakibatkan proses prediksi yang sangat lama untuk data jumlah yang sangat besar.

- 2) Algoritma K-NN tidak membedakan setiap fitur dengan suatu bobot seperti pada *Artificial Neural Network* (ANN) yang berusaha menekan fitur yang tidak berkontribusi terhadap klasifikasi menjadi 0 pada bagian bobot. K-NN tidak memiliki bobot untuk masing-masing fitur.
- 3) Karena K-NN masuk kategori *lazy learning* yang menyimpan sebagian atau semua data dan hampir tidak ada proses pelatihan, K-NN sangat cepat dalam proses pelatihan (karena memang tidak ada), tetapi sangat lambat dalam proses prediksi.
- 4) Hal yang rumit adalah menentukan K yang paling sesuai.
- 5) Karena K-NN pada prinsipnya memilih tetangga terdekat, parameter jarak juga penting untuk dipertimbangkan sesuai dengan kasus datanya. Euclidean sangat cocok untuk menggunakan jarak terdekat (lurus) antara dua data.

2.7 Fuzzy K-Nearest Neighbor

2.7.1 Konsep Fuzzy K-Nearest Neighbor

Selain K-NN yang melakukan prediksi secara tegas pada uji berdasarkan perbandingan K-tetangga terdekat, ada pendekatan lain yang melakukan prediksi juga berdasarkan K-tetangga terdekat, tapi tidak secara tegas memprediksi kelas yang harus diikuti oleh data uji, pemberian label kelas data uji pada setiap kelas dengan memberikan nilai keanggotaan seperti halnya teori himpunan *fuzzy*. Algoritma *Fuzzy K-nearest neighbor* (FK-NN) diperkenalkan oleh Keller *et al* (1985) dengan mengembangkan K-NN yang digabungkan dengan teori *fuzzy* dalam menyampaikan definisi pemberian label kelas pada data uji yang diprediksi [PRA-12].

Seperti halnya pada teori *fuzzy*, sebuah data mempunyai nilai keanggotaan pada setiap kelas, yang artinya sebuah data bisa dimiliki oleh kelas yang berbeda dengan nilai derajat keanggotaan dalam interval [0,1]. Teori himpunan *fuzzy* menggeneralisasi teori K-NN klasik dengan mendefinisikan nilai keanggotaan sebuah data pada masing-masing kelas. Formula yang digunakan:

$$u(x, c_i) = \frac{\sum_{k=1}^K u(x_k, c_k) * d(x, x_k)^{\frac{-2}{(m-1)}}}{\sum_{k=1}^K d(x, x_k)^{\frac{-2}{(m-1)}}} \quad (2.2)$$

$u(x, c_i)$ adalah nilai keanggotaan data x ke kelas c_i . K adalah jumlah tetangga terdekat yang digunakan. $u(x, c_i)$ adalah nilai keanggotaan data tetangga dalam K tetangga pada kelas c_i , nilai 1 jika data latih x_k milik kelas c_i . $d(x, x_k)$ adalah jarak dari data x ke data x_k dalam K tetangga terdekat. m adalah bobot pangkat (*weighted exponent*) yang besarnya $m > 1$.

Nilai keanggotaan suatu data pada kelas sangat dipengaruhi oleh jarak data itu ke tetangga terdekatnya. Semakin dekat ke tetangganya, semakin besar nilai keanggotaan data tersebut pada kelas tetangganya, begitu pula sebaliknya. Jarak tersebut diukur dengan N dimensi (fitur) data. Pengukuran jarak (ketidakmiripan) dua data yang digunakan dalam FK-NN digeneralisasikan dengan formula :

$$d(x_i, x_j) = (\sum_{l=1}^N |x_{il} - x_{jl}|^2)^{1/p} \quad (2.3)$$

N adalah dimensi (jumlah fitur) data, p adalah penentu jarak yang digunakan. Jika $p=1$, jarak yang digunakan adalah Manhattan. Jika $p=2$, jarak yang digunakan adalah Euclidean. Namun, jika $p = \infty$, jarak yang digunakan adalah Chebyshev [PRA-12].

2.7.2 Algoritma Fuzzy K-Nearest Neighbor

Meskipun FK-NN menggunakan nilai keanggotaan untuk menyatakan keanggotaan data pada setiap kelas, untuk memberikan keluaran akhir, FK-NN harus tetap memberikan kelas akhir hasil prediksi. Untuk keperluan ini, FK-NN memilih kelas dengan nilai keanggotaan terbesar pada data tersebut. Algoritma di bawah ini dapat memberikan detail lebih lanjut [PRA-12] :

- 1) Normalisasikan data menggunakan nilai terbesar dan terkecil data pada setiap fitur.
- 2) Cari K tetangga terdekat untuk data uji x menggunakan Persamaan 2.3.
- 3) Hitung nilai keanggotaan $u(x, c_i)$ menggunakan Persamaan 2.2 untuk setiap i , di mana $1 \leq i \leq C$.
- 4) Ambil nilai terbesar $c = u(x, c_i)$ untuk semua $1 \leq i \leq C$.
- 5) Berikan label kelas c ke data uji x .

Dengan paradigma nilai keanggotaan sebagai basis pemberian label kelas pada suatu data uji, maka suatu saat dapat terjadi ketika nilai keanggotaan agak jauh dari nilai keanggotaan penuh (1). Misalnya, ada dua kelas yang dijadikan

label, yaitu kelas 1 dan 2. Ketika ada dua data uji x dengan nilai keanggotaan 0.2 dan 0.8 berturut-turut ke kelas 1 dan 2, pemberian label kelas 2 pada data uji x rasanya masih layak karena nilai keanggotaan masih tinggi (0.8). Namun ketika nilai keanggotaan untuk kelas 1 dan 2 berturut-turut adalah 0.45 dan 0.55, rasanya agak meragukan jika data uji x diberi label kelas 2. FK-NN memungkinkan pembentukan label kelas baru (kelas belum dikenal) pada data uji yang nilai keanggotaan pada semua kelas kurang dari ambang batas (*threshold*) yang ditetapkan [PRA-12].

2.8 Akurasi

Perhitungan akurasi yang dipakai dalam penelitian ini sangat sederhana, yaitu dengan membandingkan jumlah data yang telah diklasifikasikan secara benar dengan total sampel uji, seperti terlihat pada Persamaan 2.4 berikut [NUG-06] :

$$\text{akurasi} = \frac{a}{b} * 100\% \quad (2.4)$$

Keterangan :

a = jumlah data yang telah diklasifikasikan dan bernilai benar

b = total sampel data testing yang diuji