Perbandingan *K-Nearest Neighbor* dan *Fuzzy K-Nearest Neighbor pada Diagnosis Penyakit Diabetes Melitus*

Yanita Selly Meristika[1], Achmad Ridok[2], Lailil Muflikhah[3]
[1]Mahasiswa Program Studi Informatika / Ilmu Komputer Universitas Brawijaya
[2,3]Staff Pengajar Program Studi Informatika / Ilmu Komputer Universitas Brawijaya
Program Studi Informatika / Ilmu Komuter
Program Teknologi Informasi dan Ilmu Komputer
Universitas Brawijata
Jalan Veteran Malang 65145, Indonesia
e-mail: yanitaselly7@gmail.com[1], acridokb@ub.ac.id[2] , lailil@ub.ac.id[3]

ABSTRAK

Pada perkembangan di dunia kedokteran saat ini, peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi Diabetes Melitus (DM) dan mencegah atau menghambat berkembangnya komplikasi. Hal ini dikarenakan banyaknya pasien terdiagnosis DM setelah terjadi komplikasi. Padahal DM bisa diatasi jika dideteksi lebih cepat. Salah satu metode untuk pendeteksiannya dapat menggunakan teknik data mining. Pada penelitian ini dilakukan perbandingan antara dua metode yaitu *K-Nearest Neighbor* (K-NN) dan *Fuzzy K-Nearest Neighbor* (FK-NN) untuk mendeteksi DM. *Dataset* DM diambil dari repositori UCI diabetes Indian Pima yang terdiri dari data klinis pasien terdeteksi positif dan negatif DM. K-NN merupakan teknik klasifikasi yang melakukan prediksi secara tegas pada data uji berdasarkan k tetangg terdekat. Sedangkan FK-NN melakukan prediksi data uji menggunakan nilai keanggotaan pada data uji di tiap kelas, kemudian diambil kelas dengan nilai keanggotaan terbesar dari data uji sebagai kelas hasil prediksi. Pengujian ini dilakukan terhadap 4 jumlah data latih yang berbeda yaitu 80, 130, 180, dan 230 dengan menggunakan jumlah data uji yang sama yaitu 50 data. Hasil pengujian yang dilakukan menunjukkan tingkat akurasi tertinggi terdapat pada FK-NN yakni mencapai 98%. Sedangkan K-NN akurasi tertingginya hanya mencapai 96%. Ini berarti *Fuzzy K-Nearest Neighbor* memberikan prediksi yang lebih baik dibandingkan *K-Nearest Neighbor*.

Kata kunci : Diabetes Melitus, Indian Pima, perbandingan, FKNN, KNN

I. Pendahuluan

1.1 Latar Belakang

Sejak 1991 setiap tahun, IDF (International Diabetes Federation) dan Organisasi Kesehatan Dunia (WHO) telah menetapkan tanggal 14 Nopember sebagai Hari Diabetes. Pada tahun 2007, hari diabetes ini resmi sebagai hari sedunia dalam agenda PBB. Ini menandakan kalau penyakit Diabetes Melitus (disingkat DM) tidak bisa dipandang sebelah mata. WHO memprediksi kenaikan jumlah penyandang DM di Indonesia dari 8,4 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030 [PER-11].

Pada perkembangan di dunia kedokteran saat ini, peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi DM dan mencegah atau menghambat berkembangnya komplikasi. Untuk mendeteksi seseorang terkena diabetes, ada beberapa tes lab yang harus dilakukan. US National Institute of Diabetes telah melakukan uji untuk penyakit diabetes sesuai dengan kriteria Organisasi Kesehatan Dunia yang dilakukan pada sejumlah perempuan yang berusia di atas 21 tahun, dari warisan Pima India dan tinggal di dekat Phoenix, Arizona, Amerika Serikat. Lebih dari 50% populasinya menderita DM. *Dataset* Pima meliputi delapan atribut

pengukuran dari pasien yang DM positif dan pasien didiagnosis DM negatif.

Pendeteksian penyakit diabetes menggunakan *dataset* diabetes "Indian Pima" sudah pernah dilakukan dengan berbagai metode, salah satunya implementasi metode J48 dan ID3 yang digunakan oleh [LES-12] pada data diabetes Indian prima. Akurasi tertinggi yang didapatkan pada penelitian ini mencapai 74.72% untuk J48 dan 72.64% untuk ID3.

Jika *K-Nearest Neighbor* (K-NN) melakukan prediksi dan klasifikasi secara tegas pada data uji berdasarkan perbandingan K tetangga terdekat, *fuzzy* K-NN melakukan prediksi menggunakan metode yang sama tetapi tidak secara tegas memprediksi kelas yang harus diikuti data uji. Pemberian label kelas dilakukan berdasarkan teori *fuzzy*, dimana data uji diberikan nilai keanggotaan pada setiap kelas yang terdapat pada data latih. Menurut [KEL-85], *fuzzy K-NN* adalah pengembangkan K-NN yang digabungkan dengan teori *fuzzy* dalam memberikan definisi pemberian label kelas pada data uji yang diprediksi.

Penelitian menggunakan algoritma *K-Nearest Neighbor* (K-NN) sebelumnya pernah diterapkan oleh [SHO-12] pada data diagnosis penyakit hati dan didapatkan akurasi mencapai 97.4%. Sedangkan *Fuzzy K-NN* pernah diterapkan oleh [PRA-12] pada klasifikasi bunga iris dan akurasi yang ditemukan mencapai 96%. Dapat dilihat kedua metode tersebut mempunyai kinerja yang baik dalam melakukan klasifikasi data. Namun, belum diketahui algoritma mana diantara keduanya yang lebih unggul kinerjanya. Oleh karena itu kedua algoritma ini perlu dibandingkan.

1.2. Rumusan Masalah

Rumusan masalah dalam skripsi ini adalah:

1. Bagaimana penerapan metode *K-Nearest Neighbor* dan *Fuzzy K-Nearest Neighbor* untuk klasifikasi *dataset* penyakit Diabetes Melitus (DM)?
2. Bagaiman perbandingan tingkat akurasi yang dipengaruhi sejumlah nilai data k set pada *dataset* Diabetes Melitus menggunakan aloritma *K-Nearest Neighbor* dan *Fuzzy K-Nearest Neighbor*?
3. Bagaiman pngaruh pembobotan (m) pada perhitungan *Fuzzy K-Nearest Neighbor*.

1.3 Batasan Masalah

Pada skripsi ini, permasalahan dibatasi sebagai berikut:

a. Data yang digunakan untuk penelitian dalam skripsi ini diperoleh dari database UCI machine learning repository : Indian Pima Diabetes Dataset di http://archive.ics.uci.edu
b. Paremeter DM yang digunakan berupa jumlah hamil, 2 jam PP (OGTT), tekanan diastolik, tebal kulit trisep (TSFT), 2 jam serum insulin (INS), indeks massa badan (IMB), riwayat diabetes keluarga (DPF) dan usia.
c. Output dari diagnosa resiko DM pada sistem yaitu positif DM (1) dan negatif DM (0).
d. Tidak menangani data yang memiliki *missing value* pada data latih maupun data uji.

1.4 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Menerapkan metode *K-Nearest Neighbor* dan *Fuzzy K-Nearest Neighbor* untuk klasifikasi *dataset* penyakit Diabetes Melitus (DM).
2. Membandingkan tingkat akurasi sistem yang dipengaruhi data k set pada *dataset* Diabetes Melitus menggunakan algoritma *K-Nearest Neighbor* dan *Fuzzy K-Nearest Neighbor* berdasarkan input parameter yang diberikan berdasarkan.
3. Mengetahui pengaruh dari pemberian nilai pembobotan pada proses perhitungan *Fuzzy K-Nearest Neighbor* (FK-NN).

1.5 Manfaat

Manfaat yang dapat diambil dari skripsi ini adalah dapat menemukan kinerja mana yang paling bagus antara *K-Nearest Neighbor* dan *Fuzzy K-Nearest Neighbor* dalam melakukan klasifikasi data serta penggunaan kedua algoritma ini untuk mendiagnosis penyakit Diabetes Melitus.

II. Tinjauan Pustaka

2.1. Diabetes Melitus (DM)

Diabetes Melitus merupakan penyakit metabolik dengan karakteristik hiperglikemia yang terjadi karena kelainan sekresi insulin, kerja insulin atau keduanya. jika telah terkena kronik diabetes, maka akan terjadi kerusakan jangka panjang, disfungsi atau kegagalan beberapa organ tubuh terutama mata ginjal, mata, saraf, jantung dan pembuluh darah [SET-08]. Orang yang sehat memiliki beberapa hormon insulin bertugas mengatur kadar glukosa darah. Insulin diproduksi oleh pankreas, organ kecil dekat perut yang juga mengeluarkan enzim penting yang membantu dalam proses pencernaan makanan. Insulin mengatur glukosa untuk bergerak dari darah ke dalam hati, otot dan sel lemak dimana ini digunakan untuk bahan bakar.

2.2. Data Diabetes Indian Pima

Dataset pada penelitian ini diambil dari repositori database UCI Indian Pima (http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). Dataset Indian Pima terdiri dari 768 data diagnosis DM. Data ini memiliki 8 atribut dengan target output positif diabetes (ditunjukkan dengan output 1) dan negatif diabetes (ditunjukkan dengan output 0). Daftar atribut data diabetes ditunjukkan pada table 1 berikut

Atribut	Singkatan	Deskripsi	Satuan	Type Data
Pregnant	Hamil	Banyaknya kehamilan	-	*Continous*
Plasma Glucose Concentration	OGTT	Kadar glukosa 2 jam setelah makan	mg/dl	*Continous*
Diastolic Blood Pressure	Diastolik	Tekanan darah	mmHg	*Continous*
Triceps Skin Fold Thickness	TSFT	Ketebalan kulit	mm	*Continous*
2-Hour Serum Insulin	INS	Insulin	mu U/ml	*Continous*
Body Mass Index	IMB	Berat tubuh	kg/m^2	*Continous*
Diabetes Pedigree Function	DPF	Riwayat diabetes dalam keluarga	-	*Continous*
Age	Usia	Umur pasien	Tahun	*Continous*

2.3. Logika Fuzzy

Logika *fuzzy* digunakan sebagai suatu cara untuk memetakan permasalahan dari input ke output yang diharapkan. Logika *fuzzy* bekerja menggunakan derajat keanggotaan dari sebuah nilai yang kemudian digunakan untuk menentukan hasil yang ingin dihasilkan berdasarkan atas spesifikasi yang telah ditentukan [KUS-10].

2.4. Himpunan Fuzzy

Jika di himpunan crisp, nilai keanggotaan hanya ada 0 dan 1, pada himpunan fuzzy nilai keanggotaan terletak pada rentang 0 sampai 1. Apabila x memiliki nilai keanggotaan fuzzy $\mu_A(x)=1$, artinya x menjadi anggota penuh himpunan A.

Data Mining

Data mining merupakan suatu proses yang menggunakan teknik statistika, matematika, kecerdasan buatam dan *machine learning* untuk mengidentifikasikan informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai database besar [TUR-05].

Klasifikasi merupakan proses untuk menyatakan suatu objek ke dalam salah satu ketegori yang sudah didefinisikan sebelumnya. Tujuan dari klasifikasi ini adalah *record-record* yang sebelumnya belum masuk dalam kategori dapat dinyatakan kelasnya secara akurat [KHU-07].

K-*Nearest Neighbor* merupakan salah satu algoritma yang paling sering digunakan dalam klasifikasi atau prediksi data baru. Hal ini dikarenakan algoritma ini sangat sederhana karena bekerja berdasarkan jarak terpendek dari sampel uji ke sampel latih untuk menentukan k-NN nya. Setelah mengumpulkan K-NN, lalu diambil mayoritas dari K-NN untuk dijadikan prediksi dari sampel uji. Algoritma ini mencari *k training record* (tetangga) yang memiliki jarak terdekat dari *record* baru tersebut, sehingga KNN sangat mudah diimplementasikan [SAR-00].

Langkah perhitungan K-NN dimulai dengan menghitung jarak terdekat antara tiap data uji dengan data latih. Persamaan perhitungan untuk mencari jarak menggunakan *euclidean* ditunjukkan oleh persamaan (2-2).

$$x_1 = (x_{11}, x_{12}, \ldots, x_{1n})$$

$$x_2 = (x_{21}, x_{22}, \ldots, x_{2n})$$

$$d(x_1,x_2) = \sqrt{\sum_{r=1}^{n}(a_r(x_1) - a_r(x_2))^2}$$

di mana x_1 dan x_2 merupakan dua *record* dengan n atribut. Persamaan (2-2) menghitung jarak antara x_1 dan x_2 , dengan tujuan untuk menentukan perbedaan antara nilai-nilai atribut pada record x_1 dan x_2. Setelah diketahui jarak antar *record*, kemudian diambil sebanyak *k* tetangga terdekat untuk memprediksi label kelas dari *record* baru menggunakan label kelas tetangga. Diambil kelas mayoritas sebagai kelas target output data yang baru.

Fuzzy K-Nearest Neighbor merupakan metode klasifikasi yang menggabungkan teknik *fuzzy* dengan *k-Nearest Neighbor.* Algoritma ini memberikan nilai keanggotaan kelas pada vektor sampel dan bukan menempatkan vektor pada kelas tertentu. FK-NN merupakan metode klasifikasi yang digunakan untuk memprediksi data uji menggunakan nilai derajat keanggotaan terbesar dari data uji pada setiap kelas, kemudian diambil kelas dengan nilai derajat keanggotaan terbesar dari data uji sebagai kelas hasil prediksi. Keuntungannya adalah nilai-nilai keanggotaan vektor seharusnya memberikan tingkat jaminan pada hasil klasifikasi. Formula yang digunakan ditunjukkan oleh persamaan (2-6) [Kel-85].

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(\left\|x - x_j\right\|^{\frac{-2}{(m-1)}}\right)}{\sum_{j=1}^{k}\left(\left\|x - x_j\right\|^{\frac{-2}{(m-1)}}\right)}$$

di mana:

$u_i(x)$: nilai keanggotaan data x ke kelas c_i

k : jumlah tetangga terdekat yang digunakan

u_{ij} : nilai keanggotaan kelas i pada vektor j

$x - x_j$: selisih jarak dari data x ke data x_j dalam k tetangga terdekat

m : bobot pangkat (*weight exponent*) yang besarnya m > 1

Nilai u_{ij} pada $u_i(x)$ terlebih dahulu diproses menggunakan persamaan (2-7)

$$u_{ij} = \begin{cases} 0.51 + \left(\frac{n_j}{K}\right) * 0.49 & \text{, jika } j = i \\ \left(\frac{n_j}{K}\right) * 0.49 & \text{, jika } j \neq i \end{cases}$$

di mana:

n_j : jumlah anggota kelas j pada suatu *dataset* K

K : total data latih yang digunakan

j : kelas target (baik, sedang, buruk)

Menurut [KEL-85], algoritma *fuzzy K-Nearest Neighbor* seperti berikut ini:

$W = (x_1, x_2, \ldots, x_n)$

```
BEGIN
    Input x, klasifikasi belum
diketahui
    Set K, 1 ≤ K ≤ n
    Inisialisasi i = 1
     DO UNTIL (tetangga K-terdekat x
ditemukan)
       Hitung jarak dari x ke xi
       IF (i ≤ K) THEN
        Sertakan xi di set K-tetangga
terdekat
       ELSE IF (xi lebih dekat ke x
daripada tetangga terdekat sebelumnya)
       THEN
        Hapus K-NN yang paling jauh
```

```
      Sertakan x, di set KNN
    END IF
  END DO UNTIL
Set i = 1
   DO UNTIL (x  mendapat nilai
keanggotaan di semua kelas)
    Hitung ui (x) menggunakan
(2.6)
    naikkan i
   END DO UNTIL
END
```

2.5. Akurasi

Akurasi merupakan seberapa dekat suatu angka hasil pengukuran terhadap angka sebenarnya (*true value* atau *reference value*). Tingkat akurasi diperoleh dengan perhitungan sesuai dengan persamaan (2-8) [NUG-06].

$$Akurasi = \frac{\sum data\ uji\ benar}{\sum jumlah\ total\ data\ uji} \times 100\%$$

Jumlah prediksi benar merupakan jumlah *record* data uji yang diprediksi kelasnya menggunakan metode klasifikasi dan hasilnya sama dengan kelas sebenarnya. Sedangkan jumlah total prediksi adalah jumlah keseluruhan *record* yang diprediksi kelasnya (seluruh data uji).

III.Metodologi

3.1. Langkah Penelitian

Berikut ini adalah langkah-langkah yang dilakukan dalam penelitian ini :

1. Mempelajari literatur yang memuat *K-nearest neighbor*, *fuzzy K-nearest neighbor* dan diabetes Indian Pima.
2. Mempelajari dataset yang digunakan untuk data latih , yaitu dataset diabetes Indian Pima.
3. Melakukan analisa dan perancangan sistem menggunakan algoritma *K-nearest neighbor* dan *fuzzy K-nearest neighbor*.
4. Membangun perangkat lunak berdasarkan analisa dan perancangan yang telah dilakukan (implementasi)
5. Melakukan uji coba dan mengevaluasi hasil output yang dihasilkan dari sistem.

3.2. Data Penelitian

Data yang digunakan dalam penelitian ini adalah data yang diambil dari sumber data diabetes di http://archive.ics.uci.edu. Pada *dataset* diabetes Indian Pima, terdapat 8 atribut, diantaranya jumlah hamil, 2 jam PP (OGTT), tekanan diastolik, indeks massa badan (IMB), riwayat diabetes keluarga (DPF) dan usia. Sedangkan untuk kelas output, yaitu:

0 = Negatif Diabetes Melitus (DM)
1 = Positif DM

Data diabetes Indian Pima berjumlah sebanyak 768 data klinis. Pada data ini, tidak semua atribut memiliki nilai yang lengkap dimana kelengkapan nilai atribut sangat berpengaruh pada hasil klasifikasi.

Pada penelitian ini aturan untuk mengatasi *missing value* pada masing-masing atribut sebagai berikut [LES-12]:

1. Nilai nol pada atribut hamil dapat diasumsikan bahwa nilai tersebut menyatakan pasien belum pernah melahirkan, sehingga hal ini dimungkinkan sesuai kondisi sebenarnya.
2. Data dengan nilai nol pada atribut glukosa, DBP, dan BMI dapat dihilangkan karena jumlahnya tidak terlalu banyak sehingga tidak begitu mempengaruhi hasil klasifikasi.
3. Karena atribut TSFT dan INS memiliki jumlah nilai yang tidak ada sangat besar, maka kedua atribut ini tidak mungkin dihilangkan dan tidak mungkin dipakai dalam pengklasifikasian. Oleh karena itu, dalam penelitian ini atribut TSFT dan INS tidak digunakan.

Setelah proses penanganan *missing value* dilakukan sesuai dengan aturan di atas, maka didapatkan 724 data dari 768 data aslinya. Data yang digunakan pada penelitian ini berjumlah 290 record (154 record data negatif diabetes dan 136 record data positif diabetes) untuk diproses lebih lanjut.

3.3. Deskripsi Sistem

Pada tahap awal, sistem telah memiliki data yang digunakan sebagai parameter diagnosis dimana semua nilainya berupa data numerik sehingga proses *K-nearest neighbor* dan *fuzzy K-nearest neighbor* bisa dilakukan.

Perancangan data yang dilakukan menggunakan variasi pada data latih dan data uji, di mana tiap pengujiannya menggunakan komposisi yang berubah yaitu 80, 130, 180, dan 230 data latih. sedangkan data uji yang digunakan data yang tetap yaitu 50 data.

Sistem akan melakukan proses klasifikasi pada data latih menggunakan *K-nearest neighbor* (KNN) dan *fuzzy K-nearest neighbor* (FK-NN). Tahapan dari proses ini sebagai berikut :

1. Proses input data uji dan data latih dari Microsoft excel.
2. Melakukan perhitungan normalisasi atribut menggunakan *min-max normalization*.
3. Menghitung *euclidean distance*.
4. Proses KNN yaitu mengambil mayoritas kelas pada K yang telah ditentukan sebagai kelas target pada data yang baru.
5. Proses FKNN yaitu menghitung nilai derajat keanggotaan dan mengambil nilai terbesar dari proses tersebut dan ditentukan kelas targetnya.
6. Perbandingan akurasi antara KNN dan FK-NN.

IV. Implementasi

4.1. Lingkungan Implementasi Perangkat Keras

Perangkat keras yang digunakan untuk melakukan penelitian mengenai perbandingan K-NN dan FK-NN untuk mendiagnosis diabetes melitus adalah:

1. Prosesor Intel® Core™ Duo T6600 CPU @ 2.2 GHz
2. Memori 1 GB
3. Harddisk 320 GB

4.2. Lingkungan Implementasi Perangkat Lunak

Perangkat lunak yang digunakan untuk melakukan penelitian mengenai perbandingan K-NN dan FK-NN untuk mendiagnosis diabetes melitus adalah:

1. Sistem Operasi yang digunakan yaitu Windows 7
2. Aplikasi dibangunmenggunakan bahasa pemrograman Java menggunakan Java Development Kit (JDK) 1.6 dan ditulis menggunakan editor Netbeans IDE 7.3
3. Microsoft Office Excel 2003

V. Hasil Penelitian

5.1. Hasil Pengaruh Jumlah Data Latih

Pada penelitian ini terdapat pengujian dengan 4 data latih yang berbeda yakni 80, 130, 180, dan 230 data latih dengan data uji yang sama yaitu 50 data. Dalam pengujian ini, k yang menjadi acuan adalah 13, 14, 15, dan 16.

Tabel 2 Pengaruh data Latih pada KNN

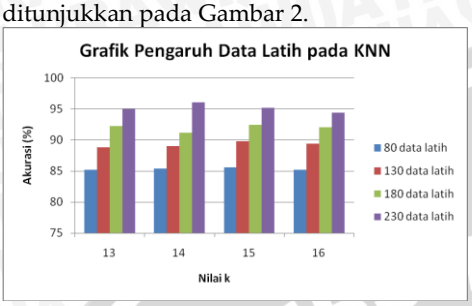
K	Akurasi Sistem (%) KNN pada			
	80 data latih	130 data latih	180 data latih	230 data latih
13	85	89	92	95
14	85	89	91	96
15	86	90	92	95
16	85	89	92	94

Tabel 3 Pengaruh data Latih pada FKNN

K	Akurasi Sistem (%) FKNN pada			
	80 data latih	130 data latih	180 data latih	230 data latih
13	84	90	94	97
14	86	90	94	97
15	86	91	95	98
16	86	91	95	97

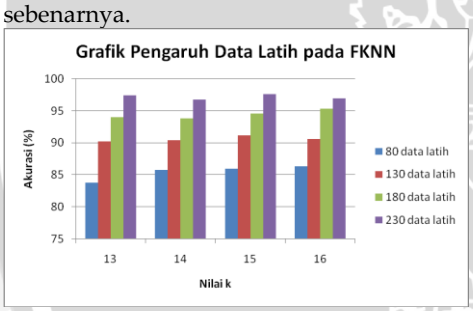
Berdasarkan hasil uji coba yang telah dilakukan, terlihat jumlah data latih sangat berpengaruh pada nilai akurasi yang dihasilkan. Pada K-NN, grafik hasil akurasi

untuk setiap jumlah data latih ditunjukkan pada Gambar 1. Sedangkan FK-NN, grafik hasil akurasi untuk setiap jumlah data latih ditunjukkan pada Gambar 2.



Gambar 1 Pengaruh Data Latih KNN

Pada gambar 1 terlihat 230 data latih memiliki akurasi yang lebih tinggi dibandingkan 80,130 dan 180 data latih. Dari grafik dapat diketahui bahwa semakin banyak jumlah data latih yang digunakan, maka hasil yang didapatkan semakin mendekati kelas prediksi sebenarnya.

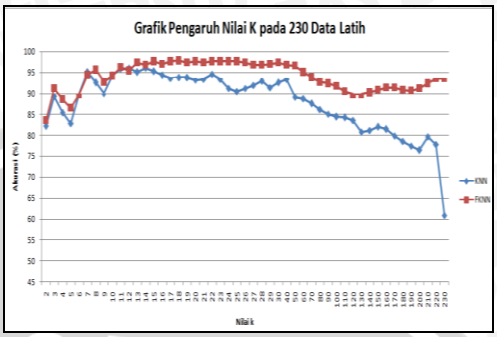


Gambar 2 Jumlah Data Latih pada FK-NN

Pada gambar 2 terlihat grafik yang dihasilkan lebih baik daripada K-NN. Dari grafik terlihat 230 data latih, akurasinya lebih bagus dibandingkan 80, 130 dan 180 data latih. Dengan meningkatnya jumlah data latih turut disertai dengan kenaikan akurasi data. Sehingga, semakin banyak data latih maka kemungkinan semakin banyak jarak *record* yang mendekati kelas data prediksi.

5.2. Hasil Pengaruh Nilai K

Pada hasil pengujian sebelumnya, data latih terbaik untuk digunakan adalah 230 data latih.

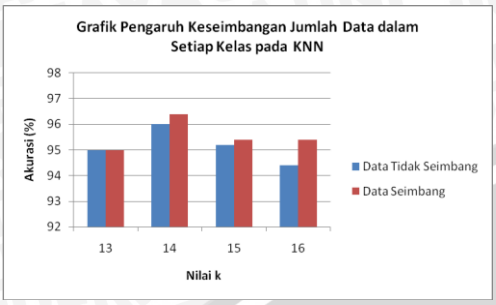


Gambar 3 Pengaruh Nilai k pada 230 Data Latih

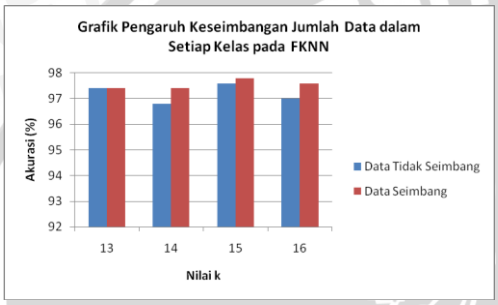
Hasil pengujian yang didapat untuk mengetahui pengaruh nilai k terhadap tingkat akurasi untuk K-NN yaitu, semakin tinggi jumlah k, maka semakin rendah akurasi yang didapatkan. Hal ini dikarenakan rentang kelas pada k yang semakin banyak memberikan nilai sensitifitas yang besar juga pada penentuan prediksi. Sedangkan pada FK-NN, walaupun juga terjadi penurunan akurasi dengan penambahan k, penurunan yang terjadi tidak signifikan seperti K-NN. Nilai akurasi FK-NN yang didapatkan cenderung stabil pada nilai k masing-masing data latih. Hal ini dikarenakan derajat keanggotaan berdasarkan jumlah kelas yang terdapat pada rentang k. Semakin tinggi derajat keanggotaannya, maka semakin tinggi pula nilai keanggotan yang dihasilkan.

5.3. Hasil Pengaruh Keseimbangan Jumlah Data dalam Setiap Kelas

Pengujian ini bertujuan untuk mengetahui pengaruh jenis *dataset* terhadap akurasi sistem. Sama dengan pengujian data latih, k yang digunakan sebagai acuan adalah k=13, 14, 15 dan 16. Dari pengujian ini ditemukan akurasi tertinggi pada KNN mencapai 96% dan pada FKNN mencapai 98%.



Gambar 4 Pengaruh keseimbangan jumlah data dalam setiap kelas pada KNN

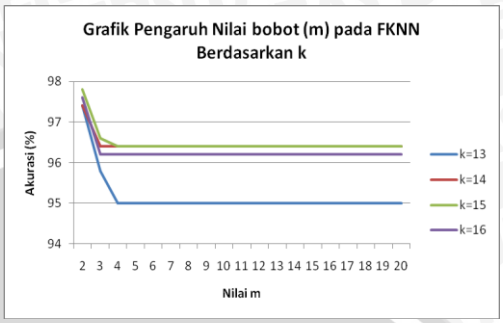


Gambar 5 Pengaruh keseimbangan jumlah data dalam setiap kelas pada FKNN

Pengujian menggunakan jenis data dengan kelas seimbang (*balanced class*) baik KNN maupun FKNN dihasilkan akurasi yang lebih meningkat dibandingkan kelas yang tidak seimbang (*imbalanced class*). Hal ini dikarenakan pada data tidak seimbang menimbulkan *noise* saat penentuan keputusan. Pada data tidak seimbang (*imbalanced class*) keputusan lebih cenderung mengacu pada data kelas yang dominan dalam *dataset*.

5.3. Hasil Pengaruh Pengaruh nilai bobot pada FK-NN

Pada pengujian ini ditemukan nilai paling optimal pada FKNN berada saat m=2. Hal ini dikarenakan m menentukan berapa banyak bobot jarak antara masing-masing tetangga ke nilai keanggotaan. Saat m mendekati 1, maka jarak berbobot semakin besar sehingga memberikan nilai maksimum pada keanggotaan kelas tersebut.



Gambar 6 Grafik Pengaruh Nilai Bobot pada FK-NN berdasarkan k

VI. Kesimpulan

Berdasarkan hasil penelitian tentang perbandingan K-Nearest Neighbor dan Fuzzy K-Nearest Neighbor pada diagnosis penyakit Diabetes Melitus (DM), dapat disimpulkan bahwa:

1. Metode K-Nearest Neighbor (K-NN) dan Fuzzy K-Nearest Neighbor (FK-NN) dapat diterapkan untuk mendiagnosis Diabetes Melitus (DM) dengan menggunakan 6 atribut yang terdapat pada data diabetes Indian Pima.
2. Nilai akurasi maksimum yang diperoleh FKNN mencapai 98% pada k=15 sedangkan KNN hanya mencapai 96% pada k=11. Ini membuktikan FKNN lebih unggul dibandingkan KNN. Tingkat akurasi pada metode KNN dan FKNN dipengaruhi oleh beberapa parameter, diantaranya:
   a. Penambahan data latih mempengaruhi peningkatan akurasi sistem, karena semakin banyaknya data latih, maka kemungkinan semakin banyak jarak *record* yang mendekati kelas data prediksi.
   b. Perubahan nilai k sangat berpengaruh terhadap akurasi sistem. Pada KNN, nilai k yang terlalu besar menyebabkan menurunnya akurasi yang didapatkan karena terdapat *noise* yang sangat besar. Semakin besar nilai K maka akan semakin banyak kemungkinan kelas diagnosis sehingga menyebabkan hasil diagnosis menjadi salah, bukan kelas yang

sebenarnya. Kebalikannya pada FKNN, nilai k yang besar membuat akurasi lebih stabil dikarenakan adanya pengaruh nilai keanggotaan tetangga terdekat. Karena nilai keanggotaan bernilai maksimum yang menjadi penentu kelas target penelitian.

c. Data latih dengan kelas yang seimbang (*balanced class*) meningkatkan akurasi dibandingkan kelas yang tidak seimbang (*imbalaced* class). Hal ini dikarenakan pada *dataset* yang tidak seimbang menimbulkan *noise* dalam penentuan keputusan. Pada data tidak seimbang (*imbalanced class*) keputusan lebih cenderung mengacu pada kelas yang mendominasi *dataset*.

3. Nilai bobot yang paling optimal pada perhitungan FKNN adalah 2. Semakin tinggi nilai m yang digunakan, maka semakin rendah nilai keanggotaan yang didapat sehingga berpengaruh pada penentuan hasil kelas prediksi**.**

VII. Saran

Saran untuk pengembangan penelitian lebih lanjut yang dapat diberikan oleh penulis adalah bisa diterapkannya penanganan nilai yang hilang pada tiap atribut untuk meningkatkan performansi sistem berikutnya.

DAFTAR PUSTAKA

[1] Han, J, Micheline, K. 2001. *Data mining Concepts and Techniques*. Morgan Kaufmann Publishers.

[2] J.Nilsson, Nill. *"Introduction To Machine Learning"*. 1996. Standford University: Standford. CA 94305

[3] Keller, James. 1985. *A Fuzzy K-Nearest Neighbor*. IEEE vol. SMC-15, No. 4

[4] Kusrinimdan Luthfi, Emha Taufiq. 2009. *Algoritma Data Mining*. Yogyakarta: Penerbit ANDI

[5] Kusumadewi, Sri, Purnomo, Hari. 2010. *Aplikasi Logika Fuzzy Untuk Pendukung Keputusan*. Yogyakarta. Graha Ilmu

[6] Lesmana, I Putu Dody. 2012. *Perbandingan Kinerja Decision Tree J48 dan ID3 Dalam Pengklasifikasian Diagnosis Penyakit Diabetes Mellitus*. JURNAL TEKNOLOGI DAN INFORMATIKA (TEKNOMATIKA). Vol. 2 No. 2

[7] Nugraha, Dany dan Ramdhany. 2006. *Diagnosis Gangguan Sistem Urinari pada Anjing dan Kucing menggunakan VFI 5.* IPB. Bandung

[8] Pramudiono Iko. 2003. *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data*. Ilmu Komputer.Com.

[9] Prasetyo, Eko. 2012. *Fuzzy K-Nearest Neighbor In Every Class Untuk Klasifikasi Data*. Seminar nasional Teknik Informatika (SANTIKA 2012). Universitas Pembangunan Nasional Veteran Jawa Timur

[10] Regina. 2012. *Penyakit Diabetes Melitus*. http://diabetesmelitus.org, diakses tanggal 23 Februari 2013

[11] Shouman, Mai; Turner, Tim; Stocker, Rob. 2012. *Applying K-Nearest Neighbor in Diagnosing Heart Disease Patients*. International Journal of Information Technology. Vol 2 No 3, June 2012.

[12] Setiawan, Meddy. 2008. *"Buku Ajar Endokrin"*. Malang. FK UMM

[13] Turban. 2005. Decision Support Systems and Intelligent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas) jilid I. Andi Offset: Yogyakarta

[14] World Health Organization Department of Noncommunicable Disease Surveillance. (1999). Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications.Geneva: Department of Noncommunicable Disease Surveillance