Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 4, April 2019, hlm. 3244-3251 http://j-ptiik.ub.ac.id

Klasifikasi Fungsi Senyawa Aktif Data Berdasarkan Kode *Simplified Molecular Input Line Entry System (SMILES)* menggunakan Metode *Modified K-Nearest Neighbor*

Yunita Dwi Alfiyanti[1], Dian Eka Ratnawati[2], Syaiful Anam[3]

[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
[3]Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Brawijaya
Email: [1]yunitadwi64@gmail.com, [2]dian_ilkom@ub.ac.id, [3]syaiful@ub.ac.id

Abstrak

Senyawa merupakan zat tunggal kimia dari dua atau lebih unsur kimia yang membentuk ikatan dan dapat diuraikan. Senyawa dibagi menjadi senyawa aktif dan senyawa tidak aktif. Senyawa aktif adalah senyawa kimia yang memiliki farmakologi atau kegunaan. Senyawa memiliki susunan yang sulit diolah pada komputer, untuk itu diciptakan kode yang mudah untuk diproses menggunakan komputer. Kode tersebut adalah *SMILES* (*Simplified Molecular Input Line Entry System*) yang adalah suatu kode ikatan kimia modern yang akan dikonversi menjadi sebuah baris sehingga memudahkan proses klasifikasi pada sistem. Karakter khusus pada *SMILES* didapat dengan melakukan *preprocessing* dengan hasil berupa 11 fitur yang terdiri dari atom B, Br, C, Cl, F, I, N, O, P, S dan OH. Fitur-fitur tersebut kemudian digunakan untuk proses klasifikasi menggunakan metode *Modified K-Nearest Neighbor*, dimana algoritme ini merupakan pengembangan dari metode KNN yang terdiri dari dua pemrosesan, validasi data latih dan pembobotan. Klasifikasi fungsi senyawa aktif bertujuan untuk mempermudah pengelompokkan senyawa aktif berdasarkan farmakologinya melalui bantuan teknologi informasi dan perosesan ilmu komputer, dimana selama ini pada bidang kedokteran memerlukan waktu yang lama dalam penentuannya karena menggunakan tes laboratorium. Pengujian yang telah dilakukan menggunakan data sebanyak 260 yang terbagi menjadi 2 kelas kategori yaitu kelas Saraf dan kelas Jantung yang terdiri dari 90% (234 data) data latih dan 10% (26 data) data uji. Pengujian tersebut mendapatkan hasil berupa nilai akurasi sebesar 73% dengan nilai k sebesar 3, sedangkan pada pengujian *k-fold cross validation* nilai akurasi didapatkan rata-rata sebesar 62,69%.

Kata kunci: *Senyawa Aktif, SMILES, Modified K-Nearest Neighbor*

Abstract

Compounds are single chemical substances from two or more chemical elements that form bonds and can be described. The compound is divided into active compounds and inactive compounds. Active compounds are chemical compounds that have pharmacology or usability. Compounds have an arrangement that is difficult to process on a computer, for which code is created that is easy to process using a computer. The code is a SMILES (Simplified Molecular Input Line Entry System) which is a code of modern chemical bonds that will be converted into a line to facilitate the classification process in the system. The special character of SMILES is obtained by doing preprocessing with the results of 11 features consisting of B, Br, C, Cl, F, I, N, O, P, S and OH atoms. These features are then used for the classification process using the Modified K-Nearest Neighbor method, where this algorithm is the development of the KNN method which consists of two processing, training data validation and weighting. The classification of the function of active compounds aims to facilitate the grouping of active compounds based on their pharmacology through the help of information technology and computer science degeneration, which so far in the medical field requires a long time in its determination because it uses laboratory tests. Tests that have been conducted using 260 data are divided into 2 categories of classes, namely the Neural class and the Heart class which consists of 90% (234 data) training data and 10% (26 data) test data. The test gets results in the form of an accuracy value of 73% with a k value of 3, whereas in the k-fold cross validation test the value of accuracy is obtained an average of 62.69%.

Keywords: *Active Compounds, SMILES, Modified K-Nearest Neighbor*

1. PENDAHULUAN

Bioinformatika merupakan ilmu berbasis multidisipliner yang menggabungkan pendekatan biologi molekuler dan teknik informatika (Searls, 2012). Untuk mewujudkan hal ini diperlukan data-data yang yang menjadi kunci penentu tindak-tanduk gejala alam tersebut, yaitu gen yang meliputi DNA (*Deoxyribose-Nucleic Acid)* atau RNA (*Ribose Nucleic Acid*). DNA atau RNA dalam ilmu biologi tersusun atas berbagai macam atom, molekul dan ion. Beberapa susunan molekul dan atom membentuk senyawa. Senyawa merupakan zat tunggal kimia yang terdiri dari dua atau lebih unsur kimia sehingga dapat membentuk ikatan serta dapat diuraikan menjadi zat yang lebih sederhana. Senyawa tersebut dapat dikategorikan menjadi senyawa aktif dan senyawa tidak aktif.

Bagi orang awam, mengetahui kegunaan dari senyawa-senyawa tersebut adalah hal yang sulit, tetapi orang dengan latar belakang ilmu pendidikan pengetahuan (Kimia dan Biologi) dapat mengerti kegunaan dari senyawa tersebut. Dari permasalahan di atas beberapa peneliti menemukan cara untuk dapat mengonversi senyawa tersebut kedalam bentuk yang mudah untuk diproses menggunakan komputer. Kode tersebut diberi nama SMILES (*Simplified Molecular Input Line Entry System*) (Weininger, 1987). Data berupa kode SMILES tersebut dapat diolah terlebih dahulu melalui proses *preprocessing* untuk mendapatkan beberapa fitur yang dibutuhkan dalam pengelolaan sistem klasifikasi kelas anti penyakit tertentu tanpa melalui tes laboratorium, namun menggunakan bantuan dari teknologi informasi dan pemrosesan ilmu komputer.

Sistem klasifikasi senyawa aktif dapat diimplementasikan menggunakan metode yang dikembangkan dalam studi Data Mining. Metode yang populer digunakan untuk klasifikasi adalah *K-Nearest Neighbor* (KNN). Algoritme ini mengklasifikasikan suatu objek baru berdasarkan kedekatan jarak suatu data dengan data yang lain. Algortime KNN memiliki kelemahan, salah satunya yaitu kelas objek baru ditentukan berdasarkan voting mayoritas kelas pada K jarak terdekat. Berdasarkan kelemahan tersebut, solusi untuk memperbaiki kinerja dari algoritme KNN dalam melakukan klasifikasi dilakukan beberapa modifikasi pada algoritme KNN yang telah diperkenalkan yaitu algoritme *Modified K-Nearest Neighbor* (MKNN). Algoritme MKNN merupakan pengembangan performansi dari metode KNN yang terdiri dari dua pemrosesan, pertama validasi data latih dan yang kedua adalah menerapkan pembobotan KNN (Parvin, Alizadeh, & Minaei-Bidgoli, 2008).

Dari penelitian yang dilakukan dengan menggunakan objek data berupa kode SMILES mendapatkan hasil akurasi dengan nilai berbeda. Penelitian Pertama yang dilakukan Ramzini, dkk (2018) dengan menerapkan metode *Learning Vector Quantization* (*LVQ*) mendapatkan hasil akurasi sebesar 76.34% dengan pembagian 80% *training set data*, *learning rate* 0,1, nilai *decrement alpha* 0,3 dan epoch 15 kali. Penelitian Selanjutnya dari Tigusti, dkk (2018) dengan mengimplementasikan metode *Fuzzy K-NN* menghasilkan akurasi sebesar 71% dengan nilai *k=15.* Penelitian berikutnya dilakukan oleh Witanto, dkk (2018) menggunakan metode K-Means dengan Inisialisasi Pusat Klaster menggunakan metode *Heuristic O(N LogN),* dari pengujian yang dilakukan penelitian ini menghasilkan akurasi sebesar 63%. Penelitian mengenai metode K-Nearest Neighbor yang digunakan oleh Leidiyana (2013) dengan objek data konsumen yang menggunakan jasa keuangan kredit kendaraan bermotor didapatkan bahwa Hasil testing untuk mengukur performa algoritme ini menggunakan metode Cross Validation, Confusion Matrix dan Kurva ROC dan menghasilkan akurasi dan nilai AUC berturut-turut 81,46 % dan 0,984 dan masih berada di tingkat baik. Begitu pula dengan penelitian yang dilakukan oleh Wafiyah, Hidayat, & Perdana (2017) menemukan bahwa melalui hasil pengujian yang dilakukan terhadap perubahan nilai *k*, perubahan jumlah data latih dan perubahan komposisi data latih mendapat akurasi rata-rata sebesar 88,55% untuk pengujian terhadapt pengaruh nilai *k*. Penelitian ini diterapkan untuk klasifikasi penyakit demam dengan mempelajari pola dari data hasil pemeriksaan sebelumnya berdasarkan 15 gejala penyakit.

Berdasarkan uraian di atas, peneliti mencoba memberikan penyelesaian klasifikasi senyawa aktif dengan menerapkan metode dalam penelitian terkait. Penelitian ini memberikan fitur dalam pengolahan data SMILES menggunakan metode MKNN untuk

mengklasifikasikan jenis farmakologi suatu senyawa dengan mudah dan sistem yang efisien.

2. DASAR TEORI

2.1. Senyawa dan Representasinya

Senyawa merupakan zat tunggal kimia yang terdiri dari dua atau lebih unsur kimia sehingga dapat membentuk ikatan serta dapat diuraikan menjadi zat yang lebih sederhana. Senyawa tersebut dapat dikategorikan menjadi senyawa aktif dan senyawa tidak aktif. Senyawa aktif adalah suatu zat yang mempunyai daya atau kemampuan melakukan pencegahan atau penyembuhan saat terjadinya berbagai macam kondisi buruk tubuh dalam proses metabolisme, senyawa aktif adalah senyawa kimia tertentu yang terdapat dalam tumbuhan dan hewan sebagai bahan obat yang mempunyai efek fisiologis terhadap organisme lain, atau sering disebut sebagai senyawa bioaktif (Salni, Marisa, & Mukti, 2011). Berbeda dengan senyawa aktif, senyawa tidak aktif tidak memiliki daya atau kemampuan pencegahan dan penyembuhan di dalam struktur atom penyusunnya.

2.1.1. *Simplified Molecular Input Line Entry System* (*SMILES*)

Simplified Molecular Input Line Entry System (SMILES) merupakan suatu cara membaca kode ikatan kimia yang dituliskan untuk melakukan pengenalan senyawa dan informasi kimia dengan cara modern. SMILES diciptakan oleh David Weininger pada tahun 1980 menggunakan konsep *graph*. Kode SMILES dituliskan dengan karakter dari kode ASCII dan disimpan dalam variabel string. Variabel dari notasi kimia tersebut lebih mudah diproses oleh komputer dan tidak membutuhkan banyak memori. Penggunaan kode SMILES yang sederhana memungkinkan pengguna mengkodekan struktur kimia yang mudah digunakan (Weininger, 1987).

2.2. *Preprocessing*

Preprocessing adalah proses untuk merubah bentuk data menjadi terstruktur sesuai dengan kebutuhan, *preprocessing* juga berguna untuk mengetahui letak dan banyak suatu huruf atau kata (*the number of terms*) (Manning, Raghavan, & Schütze, 2009). *preprocessing* dilakukan terhadap notasi *SMILES* dengan cara mencari jumlah huruf atau masing-masing unsur yang ada dalam senyawa tersebut. *Preprocessing* terhadap notasi *SMILES* adalah proses konversi notasi *SMILES* untuk mencari dan menghitung jumlah masing-masing lambang atom yang ada pada *SMILES* untuk dihitung panjang notasi *SMILES* tersebut. Panjang notasi *SMILES* dan jumlah masing-masing atom akan dijadikan sebagai masukan dalam proses klasifikasi.

2.3. K- Nearest Neighbor

Algoritme *K-Nearest Neighbor* merupakan metode pengelompokkan dalam Data Mining yang mengklasifikasikan berdasarkan data yang dengan jarak terdekat pada objek. Data diilustrasikan ke ruang dimensi dan ditampilkan fitur datanya kemudian dikumpulkan berdasarkan kelas klasifikasi data. Algoritme KNN mencari K *training record* yang memiliki jarak terdekat dari *record* baru, untuk memprediksi kelas dari *record* baru tersebut (Cahyaningtyas, Ridok, & Dewi, 2013).

Pada metode ini memiliki beberapa kelebihan seperti tangguh terhadap data pelatihan yang *noisy* dan efektif apabila data pelatihan berjumlah besar. Metode ini memiliki kekurangan juga diantaranya perlu ditentukan nilai *k* yang paling optimal yang menyatakan jumlah tetangga terdekat dan biaya komputasi yang cukup tinggi karena perhitungan jarak harus dilakukan pada setiap *query instance* secara bersama-sama dengan seluruh instan dari *training sample*.

Prinsip kerja *K-Nearest Neighbor* (KNN) adalah mencari jerak terdekat antara data yang dievaluasi dengan *k* tetangga terdekatnya dalam data pelatihan. Dekat jauhnya data dapat dihitung dengan *Euclidean distance.* Persamaan perhitungan untuk mencari jarak dengan menggunakan rumus *Euclidean distance* ditunjukkan pada Persamaan 1

$$d_{x,y} = \sqrt{\sum_{i=0}^{n}(x_{ji} - y_i)^2} \quad (1)$$

dimana,

$\boldsymbol{x} =$ Data Latih
$\boldsymbol{y} =$ Data Uji
i = Variabel Data
$d_{x,y} =$ Jarak,

tahapan perhitungan dengan metode KNN yang digambarkan dengan diagram alur seperti pada Gambar 1

Gambar 1 Proses KNN

2.3.1. Modified K-Nearest Neighbor

Modified K-Nearest Neighbor (MKNN) adalah cara untuk menemukan label kelas data baru sesuai dengan k validasi data yang sudah ditetapkan dengan perhitungan *K-Nearest Neighbor* (KNN) tertimbang (Parvin, Alizadeh, & Minaei-Bidgoli, 2008). Dalam algoritme MKNN, setiap data pada data *training* harus melalui proses validitas dulu. Nilai ini tergantung pada setiap nilai tetangganya. Tujuan utama yang menjadi dasar modifikasi pada metode *KNN* ini adalah menentukan kelas label dari *query instance* ke dalam K data latih yang telah divalidasi. Setelah menentukan *Validitas*, *weighted* KNN dilakukan pada setiap data uji.

Secara garis besar terdapat dua proses utama dalam metode ini, yaitu (Parvin, Alizadeh, & Minaei-Bidgoli, 2008):

1. *Validitas* data latih

Proses *validitas* dilakukan untuk semua data pada data *training*. *Validitas* setiap data tergantung pada setiap tetangganya. Setelah dihitung *validitas* tiap data maka nilai v*aliditas* tersebut digunakan sebagai informasi lebih mengenai data tersebut. Untuk menghitung v*aliditas* dari data pada data *training*, tetangga terdekatnya perlu dipertimbangkan. Di antara tetangga terdekat dengan data, *validitas* digunakan untuk menghitung jumlah titik dengan label yang sama untuk data tersebut. Persamaan yang digunakan untuk menghitung *validitas* dari setiap titik pada data *training* adalah seperti pada Persamaan 2

$$Validitas(x) = \frac{1}{K}\sum_{i-1}^{K} S(lbl(x), (lbl(N_i(x)))) \quad (2)$$

dimana:

K = Jumlah titik terdekat

$lbl(x)$ = Kelas x

$N_i(x)$ = Label kelas titik terdekat x,

fungsi S pada Persamaan 2 digunakan untuk menghitung kesamaan antara titik x dan data ke-I dari tetangga terdekat. Fungsi S dituliskan dengan Persamaan 3

$$S(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} \quad (3)$$

dimana,

a = Kelas a pada data training

b = Kelas lain selain kelas a pada data training, melalui Persamaan 3 dapat diketahui bahwa a dan b adalah label kelas kategori suatu data latih. S bernilai 1 jika label kategori a sama dengan kabel kategori b, sedangkan S bernilai 0 jika label kategori a tidak sama dengan label kategori b.

2. *Weight voting*

Weight voting adalah salah satu variasi metode *KNN* yang menggunakan K tetangga terdekat dan hasil perhitungan dari jarak masing-masing data. Pada metode *MKNN*, masing-masing K tetangga terdekat dihitung menggunakan Persamaan 2 dan Persamaan 3. Nilai *validitas* yang dihasilkan dari setiap data yang dihitung sebelumnya kemudikan dikaliakan dengan hasil *weight voting* berdasarkan jarak. Sehingga dalam metode *MKNN*, nilai rumus persamaan untuk menghitung *weight voting* dinyatakan pada Persamaan 4

$$W(x) = Validitas(x)\frac{1}{d(x,y)+0{,}5} \quad (4)$$

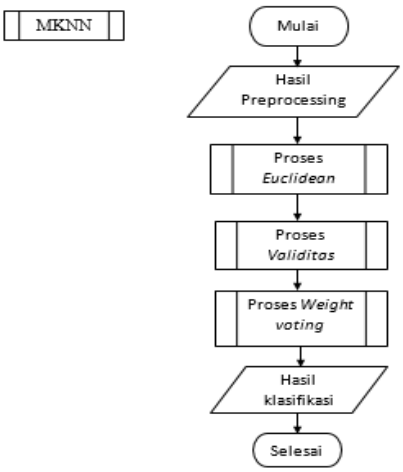
dimana:

$W(x)$ = Nilai *Weight voting*

$d(x,y)$ = Jarak *Euclidean*

$0{,}5$ = Konstanta.

Secara umum alur dari proses klasifikasi menggunakan metode *Modified K-Nearest Neighbor* ditunjukkan pada Gambar 2



Gambar 2 Alur proses perhitungan menggunakan metode *Modified k-nearest neighbor*

Pada penelitian ini proses klasifikasi dengan MKNN menggunakan 2 kelas kategori. Proses MKNN dijabarkan sebagai berikut:

1. Memasukkan data kode SMILES berupa jumlah masing-masing elemen beserta panjang kode SMILES.
2. Menginisialisasi nilai K tetangga.
3. Melakukan perhitungan jarak untuk setiap data latih menggunakan rumus Persamaan 1 yaitu *eucledian distance*.
4. Melakukan perhitungan nilai *Validitas* setiap data latih menggunakan rumus Persamaan 2.
5. Melakukan perhitungan jarak untuk setiap data uji dengan data latih menggunakan rumus Persamaan 1.
6. Melakukan perhitungan *weight voting* pada setiap data uji menggunakan rumus Persamaan 4
7. Menentukan kelas klasifikasi dari data uji sesuai nilai *weight voting* terbesar.

3. METODOLOGI

Metodologi yang digunakan untuk penelitian klasifikasi senyawa aktif menggunakan metode *Modified K- Nearest Neighbor*. Dalam metodologi penelitian ini juga menjelaskan tentang studi kepustakaan, teknik pengumpulan data yang digunakan dalam penelitian, lokasi penelitian, strategi dalam penelitian, tipe penelitian, implementasi algoritme *M-KNN* dan tentang teknik analisis data.

Penelitian dimulai dengan studi kepustakaan yang bertujuan untuk mempelajari mengenai metode yang digunakan. Studi dilakukan dengan meggunakan penelitian sebelumnya dengan menggunakan metode yang sama. Setelah melakukan studi literatur maka langkah selanjutnya adalah menentukan tipe penelitian dan merancang strategi penelitian agar pelaksanaan penelitian lebih tersusun. Kemudian menentukan lokasi makna yang akan digunakan dalam penelitian dan dilanjutkan dengan pengumpulan data. Data yang digunakan pada penelitian ini berupa notasi *SMILES* yang didapatkan dari halaman resmi *pubchem*. Langkah selanjutnya yaitu analisis kebutuhan yang dimaksudkan untuk mengetahui kebutuhan apa yang akan digunakan pada penelitian ini. Selanjutnya yaitu perancangan dan implementasi pada sistem. Setelah sistem berhasil diimplementasikan maka selanjutnya dilakukan pengujian dan analisis mengenai hasil keluaran. Kemudian terakhir adalah penarikan kesimpulan dan saran yang diberikan penulis dari penelitian yang telah dilakukan.

4. HASIL

4.1. Pengujian Variasi Nilai *k*

Pengujian ini menggunakan *k* yang berbeda yang dilakukan sebanyak 5 kali dengan nilai *k* sebesar *k* = 2, *k* = 3, *k* = 5, *k* =7 dan *k* =9. Tujuan dari penelitian ini adalah untuk mengetahui pengaruh nilai tetangga terdekat terhadap nilai akurasi. Dalam scenario pengujian ini digunakan data sebanyak 260 datayang terbagi menjadi latih dan data uji dengan presentase data sebesar 90% (234 data) data latih dan 10% (26 data) data uji. Hasil dari pengujian ditunjukkan pada Tabel 1 dan Gambar 3.

Tabel 1 Hasil Pengujian Variasi Nilai *k*

Skenario Ke-	Nilai *k*	Nilai Akurasi
1	2	69%
2	3	73%
3	5	69%
4	7	69%
5	9	58%

Gambar 3 Grafik hasil pengujian variasi nilai *k*

Pada Tabel 1 dan Gambar 3 diketahui akurasi terbesar pada nilai *k*=3 sebesar 73%. Melihat dari grafik Gambar 4 didapatkan hasil bahwa setiap pengujian dengan menggunakan nilai *k* yang berbeda menghasilkan nilai akurasi yang berbeda. Nilai *k* dapat mempengaruhi besarnya nilai akurasi pada pengujian dikarenakan jika nilai *k* semakin besar maka semakin banyak data yang tidak relevan diikutkan dalam pengambilan keputusan hasil klasifikasi sehingga nilai akurasi menurun.

4.2 Pengujian *Holdout Validation* Terhadap Data Latih

Pengujian *holdout validation* terhadap data latih digunakan untuk mengetahui apakah jumlah data latih dapat mempengaruhi nilai akurasi. Dalam Metode *holdout*, data awal yang diberi label dibagi ke dalam dua himpunan secara random yang dinamakan data latih dan data uji dengan jumlah keseluruhan data sebesar 100%. Pada pengujian ini digunakan seluruh data yaitu 260 data dan memiliki presentase jumlah yang berbeda-beda dan terdapat 4 pengujian dengan jumlah data latih dan data uji yang berbeda, dimana dalam scenario pengujian ini menggunakan nilai *k* dari pengujian sebelumnya yaitu *k*=3. Jumlah masing-masing data untuk pengujian ini dapat dilihat pada Tabel 2 dan Gambar 4.

Tabel 2 Hasil pengujian *holdout validation*

Skenario Ke-	Persentase Jumlah Data Latih	Presentase Jumlah Data Uji	Nilai akurasi
1	60%	40%	68%
2	70%	30%	72%
3	80%	20%	71%
4	90%	10%	73%

Gambar 4 Grafik hasil pengujian *holdout validation*

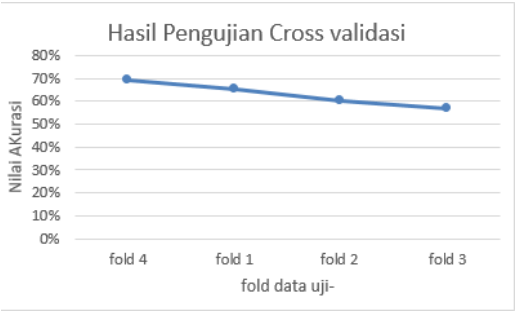
Pada Tabel 2 dan Gambar 4 mendapat hasil bahwa pada pengujian dengan menggunakan data latih 90% dan data uji sebesar 10% tersebut mendapatkan hasil akurasi mencapai nilai 73%. Berdasarkan penjelasan tersebut dapat disimpulkan bahwa secara umum jika semakin banyak jumlah data latih yang digunakan pada proses klasifikasi maka nilai akurasi semakin tinggi karena sistem melakukan proses pembelajaran lebih banyak.

4.3 Pengujian *K-Fold Cross Validation*

K-Fold Cross validation adalah pengujian untuk mengetahui kelayakan seluruh data dengan sistem acak. Dalam metode k-fold, dataset yang dibagi menjadi sejumlah k-buah partisi secara acak ke dalam k partisi yang berukuran sama.dan dilakukan sejumlah k-kali eksperimen. Pengujian ini dilakukan dengan melakukan pembagian dataset menjadi 4 kelompok (4 fold) dengan banyak data yang sama yaitu 65 data untuk masing-masing kelompok, kemudian setiap kelompok akan dibagi menjadi data latih dan data uji secara bergantian dan diujikan menggunakan nilai *k*=3. Skenario pengujian ini terbagi menjadi 4 yang hasilnya dapat dilihat pada Tabel 3 dan Gambar 5.

Tabel 3 Hasil pengujian *k-fold cross validation*

Skenario Ke-	fold Data Latih	fold Data Uji	Nilai akurasi
1	1,2dan 3	4	69%
2	2,3 dan 4	1	65%
3	1,3 dan 4	2	60%
4	1,2 dan 4	3	57%



Gambar 5 Grafik hasil pengujian *k-fold cross validation*

Pada Gambar 5 dan Tabel 3 pengujian tersebut menghasilkan akurasi tertinggi sebesar 69% pada pengujian fold data uji ke-4. Berdasarkan penjelasan tersebut dapat disimpulkan bahwa penggunaan data yang acak pada klasifikasi menunjukkan bahwa setiap data memiliki kesempatan untuk dijadikan data latih maupun data uji dikarenakan setiap kelompok data memiliki karakteristik data yang berbeda.

5. KESIMPULAN

1. Cara melakukakan *preprocessing* terhadap data SMILES adalah mencari atau menemukan karakter dari masing-masing SMILES kemudian menghitung panjang

notasi dan jumlah masing-masing atom penyusunnya. Langkah Selanjutnya adalah membagi setiap jumlah atom yang sudah diketahui dengan panjang notasi SMILES sehingga dapat dijadikan fitur untuk proses klasifikasi. Preprocessing dari *SMILES* tersebut mendapatkani fitur berupa jumlah atom *B, C, N, O, P, S, F, Cl, Br, I, dan OH*.

2. Cara kerja dari metode *Modified k-nearest neighbor* yaitu menghitung selisih nilai dari masing-masing data latih atau disebut *Euclidean distance*. Kemudian nilai *Euclidean distance* diurutkan dari nilai terendah ke nilai tertinggi. Setelah nilai *Euclidean distance* diurutkan data diambil sebanyak *k* dan dihitung nilai *validitas* dari setiap data latih tersebut. Perhitungan *Euclidean distance* juga dilakukan pada data latih dengan data uji dan diurutkan untuk diambil nilai tertinggi, kemudian dari proses tersebut dilakukan perhitungan weigh voting untuk mendapatkan kelas klasifikasi berdasarkan nilai bobot kelas tertingginya. Kelas klasifikasi yang digunakan pada penelitian ini yaitu Saraf dan Jantung.
3. Pada penelitian ini metode *Modified k-nearest neighbor* memberikan hasil akurasi yang cukup baik dalam melakukan proses klasifikasi fungsi senyawa menggunakan kode *SMILES*. Setelah melakukan beberapa pengujian hasil yang didapatkan adalah sebagai berikut:
    a. Pengujian validasi program untuk memastikan hasil perhitungan manualisasi dan hasil keluaran sistem memberikan hasil yang valid.
    b. Pengujian *k-fold cross validation* sebanyak 4 kali percobaan menghasilkan nilai akurasi tertinggi sebesar 69% dengan rata-rata akurasi dari seluruh percobaan sebesar 62.69%.
    c. Pengujian variasi nilai *k* menghasilkan nilai akurasi tertinggi sebesar 73% dengan nilai K=3.
    d. Pengujian *holdout validation* terhadap jumlah data latih menghasilkan nilai akurasi tertinggi sebesar 73% dengan menggunakan 90% data latih dan 10% data uji.

6. DAFTAR PUSTAKA

Cahyaningtyas, Y., Ridok, A., & Dewi, C. (2013). Penerapan Fuzzy K-Nearest Neighbor untuk Menentukan Status Evaluasi Kinerja Karyawan. *Repositori Jurnal Mahasiswa PTIIK UB, 1*(4)

Kurniawan, E. (2017). Analisa Data Rekam Medis Menggunakan Teknik Data Mining Association Rules Dengan Algoritme Clustering. *Seminar Nasional Pendidikan, Sains dan Teknologi* .

Leidiyana, H. (2013). Penerapan Algoritme K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor. *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic, 1*(1), 65-76.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* England: Cambridge University Press.

Muliantara, A. (2009). Penerapan Regular Expression Dalam Melindungi Alamat Email Dari Spam Robot Pada Konten Wordpress. *Jurnal Ilmu Komputer, 2*(1).

Parvin, H., Alizadeh, H., & Minaei-Bidgoli, B. (2008). MKNN: Modified K-Nearest Neighbor. *Proceedings of the World Congress on Engineering and Computer Science 2008.*

Ramzini, S., Ratnawati, D. E., & Anam, S. (2018). Penerapan Metode *Learning Vector Quantization* (*LVQ*) untuk Klasifikasi Fungsi Senyawa Aktif Menggunakan Notasi *Simplified Molecular Input Line System* (*SMILES*). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(12), 6160-6168.

Salni, Marisa, H., & Mukti, R. W. (2011). Isolasi Senyawa Antibakteri dari Daun Jengkol (Pithecolobium Lobatum Benth) dan Penentuan Nilai KHM-nya. *Jurnal Penelitian Sains, 14*(1(D)), 38-41.

Searls, D. B. (2012). A New Online Computational Biology Curriculum. *PLoS Comput Biol, 10*(6). doi:e1003662

Tigusti, R. R. W., Ratnawati, D. E., & Anam, S. (2018). Implementasi *Fuzzy K-Nearest Neighbor (FK-NN)* Untuk Mengklasifikasi Fungsi Senyawa Berdasarkan *Simplified Molecular Input Line Entry System (SMILES)*. *Jurnal*

Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(12), 6331-6338.

Wafiyah, F., Hidayat, N., & Perdana, R. S. (2017). Implementasi Algoritme Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Demam. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 1*(10), 1210-1219 .

Weininger, D. (1987). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci, 28*(1), 31-36.

Witanto, S., Ratnawati, D. E., & Anam, S. (2018). Pengelompokan Fungsi Aktif Senyawa Data SMILES (*Simplified Molecular Input Line Entry System*) Menggunakan Metode K-Means Dengan Inisialisasi Pusat Klaster Menggunakan Metode *Heuristic O(N LogN)*. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(1), 8291-8296.