DOI: 10.22146/jnteti.v13i2.9579

Optimasi Algoritma *K-Nearest Neighbors* Berdasarkan Perbandingan Analisis *Outlier* (Berbasis Jarak, Kepadatan, LOF)

Fitri Ayuning Tyas[1], Mahda Nurayuni[1], Hidayatur Rakhmawati[1]

[1] Program Studi Sistem Informasi, STMIK Muhammadiyah Paguyangan Brebes, Brebes, Indonesia

[Diserahkan: 19 September 2023, Direvisi: 1 Desember 2023, Diterima: 20 Maret 2024]
Penulis Korespondensi: Fitri Ayuning Tyas (email: tyas_fa@stmikmpb.ac.id)

INTISARI — Pertumbuhan data yang terjadi saat ini berpengaruh terhadap analisis data di berbagai bidang, seperti astronomi, bisnis, kedokteran, pendidikan, dan finansial. Data yang terkumpul dan tersimpan mengandung nilai ekstrem atau nilai pengamatan yang berbeda dari kebanyakan nilai hasil pengamatan lain. Nilai ekstrem tersebut disebut dengan *outlier*. *Outlier* pada sebagian data sering kali memiliki nilai yang mengandung informasi penting, sehingga perlu dikaji agar dapat diambil keputusan untuk menghapus atau menggunakan data tersebut sebelum diterapkan dalam penambangan data. Deteksi *outlier* dapat dilakukan sebagai prapemrosesan data dengan menggunakan teknik analisis *outlier*. Beberapa teknik analisis *outlier* yang banyak diterapkan antara lain metode berbasis jarak (*distance*), metode berbasis kepadatan (*density*), dan metode *local outlier factor* (LOF). *K-nearest neighbors* (KNN) merupakan salah satu algoritma penambangan data yang sangat sensitif terhadap *outlier* karena cara kerjanya yang bergantung pada nilai *k*. Oleh karena itu, perlu penanganan tepat saat KNN bekerja pada *dataset* yang mengandung *outlier*. Metode eksperimen dipilih dalam menerapkan metode usulan, dengan tujuan untuk mengoptimasi algoritma KNN berdasarkan perbandingan analisis *outlier* (KNN-*distance*, KNN-density, dan KNN-LOF). Hasil penelitian menunjukkan bahwa KNN-kepadatan unggul sebanyak tiga kali: pada Wisconsin Breast Cancer dengan nilai rata-rata akurasi sebesar 99,34% pada *k*=3 dan *k*=5; pada *Glass* dengan nilai rata-rata akurasi sebesar 85,25% pada *k*=7; dan pada *Lymphography* dengan nilai rata-rata akurasi sebesar 85,45% pada *k*=5. Selanjutnya, berdasarkan hasil uji Friedman dan uji Nemenyi, juga terbukti bahwa ada perbedaan yang signifikan antara KNN-kepadatan dengan KNN-LOF.

KATA KUNCI — *K-Nearest Neighbors*, *Outlier*, Kepadatan, Jarak, LOF, Uji Friedman, Uji Nemenyi.

I. PENDAHULUAN

Pertumbuhan data dari segi jumlah produksi maupun data yang disimpan mengalami peningkatan setiap waktu, tetapi data-data tersebut belum bernilai sebagai informasi. Analisis data merupakan kebutuhan penting agar data dapat diubah menjadi informasi, sehingga dapat digunakan di berbagai bidang, seperti astronomi, bisnis, kedokteran, pendidikan, dan finansial [1], [2]. Dengan penambangan data (*data mining*), data dapat digunakan untuk memenuhi kebutuhan sebagai alat dalam menemukan pengetahuan dari data [1], [3]. Pengetahuan dari data dapat berupa pola, rumus, pohon keputusan, dan yang lainnya. Penambangan data adalah studi tentang mengumpulkan, membersihkan, memproses, menganalisis data yang sudah ada, dan memperoleh manfaat wawasan dari data [4], [5]. Maka, data yang semula hanya berisi kumpulan fakta akan bernilai sebagai pengetahuan atau memiliki pola setelah diolah dengan penambangan data.

Outlier disebut sebagai kelainan, sumbang, menyimpang, atau anomali dalam penambangan data dan literatur statistik [5]. Dalam perspektif klasifikasi, *outlier* sering direpresentasikan sebagai bentuk fitur yang tidak signifikan, nilai yang hilang, atau *instance* yang berlebihan atau tidak konsisten [6]. *Outlier* memiliki dampak yang buruk terhadap hasil analisis data, sehingga *outlier* harus mendapatkan penanganan secara khusus [7]. Walaupun memiliki perilaku berbeda dengan mayoritas data yang lain dan sering dianggap sebagai derau (*noise*), *outlier* sering kali mengandung informasi yang berguna [8]. Derau adalah varian acak dalam variabel terukur yang dapat diartikan sebagai penyimpangan nilai atribut atau nilai yang hilang (*missing value*) atau nilai yang salah dan termasuk sebagai *outlier* [1], [9]. *Outlier* berbeda dari derau [1], tetapi derau menjadi bagian dari *outlier*. Ekstraksi pengetahuan dari data yang mengandung derau atau *outlier* adalah masalah kompleks di bidang penambangan data [10]. Permasalahan deteksi *outlier* masih terus dikembangkan dalam beberapa penelitian.

Deteksi *outlier* penting dilakukan dalam prapemrosesan data karena jika terdapat *outlier* saat penerapan penambangan data, kemungkinan akan dihasilkan keluaran yang tidak akurat [11]. Prapemrosesan data adalah teknik penanganan data sebelum masuk pada tahap pengolahan data, seperti pembersihan data, transformasi data, dan standardisasi data [1], [12]. Deteksi *outlier* mempunyai peran besar, seperti pengambilan keputusan, pengelompokan, dan klasifikasi pola, karena dapat mengungkapkan fenomena yang jarang tetapi penting serta menemukan pola yang menarik atau tidak terduga [13]. Teknik deteksi *outlier* dibagi menjadi metode berbasis statistika, klaster, jarak (*distance-based*), dan kepadatan (*density-based*) [14], [15]. RapidMiner merupakan platform penambangan data yang digunakan untuk mengolah data, salah satunya menyediakan fasilitas deteksi *outlier*, seperti deteksi berbasis jarak, kepadatan, dan *local outlier factor* (LOF). Dengan demikian, dibutuhkan pemilihan deteksi *outlier* yang tepat untuk mendeteksi keberadaan *outlier* pada data.

K-nearest neighbors (KNN) merupakan salah satu algoritma pembalajaran malas (*lazy learning algorithm*) yang populer sebagai algoritma klasifikasi penambangan data karena memiliki konsep sederhana, mudah diimplementasikan,

memiliki kinerja yang baik, dan sukses dikembangkan di berbagai aplikasi [16]–[18]. KNN bekerja dengan menemukan jarak terdekat antara sejumlah *k* objek data atau pola pada data latih dan data uji, kemudian memilih kelas berdasarkan jumlah pola terbanyak di antara *k* pola tersebut [16]. Penentuan nilai *k* menjadi masalah penting pada KNN, khususnya di bidang deteksi *outlier*. Jika nilai *k* terlalu kecil, hasilnya akan sensitif terhadap *outlier*. Sebaliknya, jika nilai *k* terlalu besar, hasilnya akan lebih resistan terhadap *outlier* [6], [16], [19]. Dengan demikian, perlu adanya penanganan yang tepat pada KNN saat bekerja dengan *dataset* yang memiliki *outlier*.

Penelitian ini mengusulkan optimasi algoritma KNN berdasarkan analisis *outlier* (berbasis jarak, kepadatan, dan LOF). Metode usulan tersebut diberi nama KNN-*distance*, KNN-*density*, dan KNN-LOF. Tujuan dari deteksi *outlier* adalah mengetahui nilai atau jumlah *outlier* pada *dataset*. Nilai yang mengandung *outlier* akan dihapus dari *dataset*. Selanjutnya, evaluasi kinerja model metode KNN-*distance*, KNN-*density*, dan KNN-LOF diukur menggunakan 10-*fold cross validation*. Rata-rata akurasi yang dihasilkan dibandingkan berdasarkan tingkat signifikasi menggunakan uji Friedman. Hasil uji Friedman digunakan untuk menunjukkan adanya perbedaan antara metode-metode yang diusulkan, sedangkan uji Nemenyi dilakukan untuk mengetahui metode usulan yang memiliki perbedaan paling signifikan.

Bagian selanjutnya dari makalah ini menjelaskan beberapa metode yang digunakan, metodologi penelitian, hingga hasil yang didapatkan. Pada Bagian II dijelaskan beberapa landasan teori mengenai deteksi *outlier*. Bagian III membahas metodologi penelitian yang telah dilakukan, mulai dari pengumpulan *dataset*, eksperimen metode usulan, dan evaluasi hasil eksperimen. Bagian IV membahas hasil dari pengujian yang telah dilakukan dan Bagian V menyajikan kesimpulan berdasarkan hasil pengujian.

II. DETEKSI *OUTLIER*

Deteksi *outlier* merupakan langkah utama untuk menerapkan penambangan data [20] dan banyak digunakan pada penelitian deteksi kasus-kasus abnormal basis data [21]. Pada pembahasan sebelumnya, disebutkan bahwa di antara teknik deteksi *outlier* adalah metode berbasis statistika, klaster, jarak , dan kepadatan. Metode berbasis statistika adalah metode pendeteksian *outlier* dengan cara menghitung nilai rata-rata titik data. Contoh metode berbasis statistika adalah distribusi Gaussian dan histrogram [2], [13], [22]. Titik-titik dengan probabilitas rendah yang dihasilkan oleh metode distribusi dianggap sebagai *outlier* [22]. Metode berbasis klaster adalah pengelompokan objek yang sejenis dengan menghitung matriks jarak antarklaster. Contoh metode berbasis klaster adalah *K-means*, *self-organizing map* (SOM), dan *one class support vector machine* (SVM) [13], [22], [23]. Titik data yang jauh dari klaster atau kelompoknya dianggap sebagai *outlier*. Metode berbasis jarak merupakan pendekatan berbasis jarak yang bekerja dengan menghitung jarak suatu titik data dari tetangganya. Objek yang jaraknya lebih jauh dari tetangganya disebut *outlier* [22]. Contoh metode berbasis jarak antara lain KNN dan *outlier detection using indegree number* (ODIN) [13], [23], [24]. Kepadatan suatu titik data dihitung dan dibandingkan dengan titik data di sekitarnya dan ini disebut sebagai skor *outlier* dalam metode berbasis kepadatan [22]. Seharusnya titik data normal dan titik data tetangga mempunyai kepadatan yang sama. Di sisi lain, *outlier* mempunyai kepadatan yang berbeda [22]. Metode berbasis kepadatan diusulkan untuk mengatasi kekurangan deteksi *outlier* global berbasis jarak. Contoh metode berbasis kepadatan adalah *local outlier probabilities* (LoOP), *local correlation integral* (LOCI), dan LOF [2], [9], [13], [15], [22]. Metode LOF menjadi metode deteksi *outlier* berbasis kepadatan yang sangat populer [2]. LOF bekerja dengan cara menguji rasio kepadatan lokal di sekitar suatu objek dengan kepadatan rata-rata yang dapat dijangkau dari objek-objek di sekitarnya. Lingkungan dari objek tersebut ditentukan berdasarkan parameter *k* tetangga minimum yang diberikan oleh pengguna dan jarak tetangga terdekat [15], [23]. Dari perbedaan metode pendeteksi *outlier* di atas, dibutuhkan pemilihan metode yang tepat untuk mengoptimasi kinerja sebuah algoritma penambangan data.

TABEL I
KUMPULAN DATASET EKSPERIMEN

No	Nama *Dataset*	Jumlah *Instance*	Atribut dan Label
1.	Wisconsin Breast Cancer	699	9 atribut + 1 label
2.	Glass	214	10 atribut + 1 label
3.	Harbeman	306	3 atribut + 1 label
4.	Lymphography	148	18 atribut + 1 label
5.	Parkinsons	195	22 atribut + 1 label

III. METODOLOGI

Metode yang digunakan dalam penelitian ini adalah metode eksperimen dengan tahapan pengumpulan *dataset*, eksperimen penerapan metode usulan (KNN-*distance*, KNN-*density* dan KNN-LOF), dan evaluasi hasil eksperimen.

A. PENGUMPULAN DATASET

Dataset yang digunakan pada penelitian ini adalah *dataset* yang mengandung *outlier*. *Dataset* tersebut bersifat publik dan dapat diunduh di UCI *Machine Learning Repository* https://archive.ics.uci.edu/ml/index.php serta Kaggle https://www.kaggle.com/ [3], [13].

Tabel I menunjukkan rincian lima *dataset* yang digunakan dalam eksperimen penerapan metode usulan. *Dataset* tersebut adalah Wisconsin Breast Cancer, Glass, Haberman, Lymphography, dan Parkinson.

B. EKSPERIMEN METODE USULAN

Metode yang diusulkan mengoptimasi kinerja algoritma KNN berdasarkan perbandingan analisis *outlier* (KNN-*distance*, KNN-*density*, dan KNN-LOF), yang dilakukan menggunakan RapidMiner. Langkah-langkah eksperimen metode usulan diperlihatkan pada Gambar 1.

1) MASUKAN *DATASET*

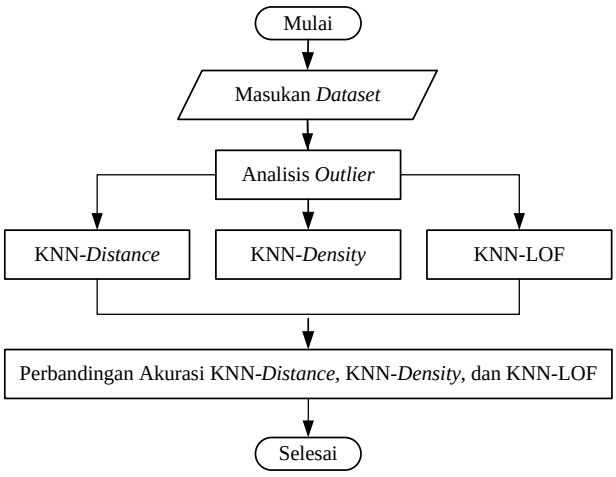
Langkah pertama yang dilakukan dalam eksperimen ini adalah mengimpor *dataset* yang sebelumnya telah diperoleh dari proses pengumpulan data.

2) ANALISIS *OUTLIER*

Teknik deteksi *outlier* yang diterapkan adalah KNN-*distance*, KNN-*density*, dan KNN-LOF. Deteksi *outlier* merupakan bagian dari prapemrosesan data. Tujuan deteksi *outlier* dalam tahapan ini adalah mencari kumpulan data yang termasuk dalam *outlier*, sehingga jumlah *outlier* pada *dataset* dapat diketahui.

3) PERBANDINGAN AKURASI METODE USULAN

Hasil pemodelan dari metode analisis *outlier* berupa rata-rata akurasi. Hasil rata-rata akurasi tersebut selanjutnya dibandingkan tingkat signifikasinya menggunakan uji



Gambar 1. Langkah-langkah eksperimen.

Friedman dan uji Nemenyi. Uji Friedman adalah analisis nonparametrik untuk pengujian analisis variasi dua arah berdasarkan peringkat yang diusulkan oleh Demsar [25]. Dalam penelitian ini, uji Friedman digunakan untuk membandingkan kinerja metode usulan berdasarkan *chi*-kuadrat (*chi-square*) atau *F-disrtibution* atau *P-value.*

Gambar 2 menjelaskan langkah-langkah perbandingan kinerja metode usulan (KNN-*distance*, KNN-*density*, dan KNN-LOF). Langkah awal adalah menyiapkan data pengamatan dari hasil eksperimen dan melakukan pemeringkatan. Hipotesis yang ditetapkan pada uji Friedman dalam penelitian ini adalah sebagai berikut.

- H_0 (hipotesis nol): tidak ada perbedaan nilai rata-rata akurasi antarmetode usulan yang digunakan pada eksperimen dalam penelitian ini.
- H_a (hipotesis alternatif): ada perbedaan nilai rata-rata akurasi antarmetode yang digunakan pada eksperimen dalam penelitian ini.

Nilai taraf signifikansi (α), atau disebut dengan tingkat kesalahan, ditetapkan untuk pengambilan keputusan dari pengujian hipotesis. Nilai taraf signifikansi yang dapat digunakan adalah 0,05 (5%) dan 0,1 (10%). Makin kecil nilainya, makin besar tingkat kepercayaan pengambil keputusan.

Langkah selanjutnya adalah menghitung statistik uji Friedman. Pada penelitian ini, perhitungan statistik uji Friedman dilakukan menggunakan *chi*-kuadrat dan *F-distribution*.

Uji *chi*-kuadrat adalah salah satu jenis uji komparatif nonparametrik yang dilakukan pada dua variabel, dengan skala data kedua variabel adalah nominal [26], [27]. Persamaan (1) digunakan pada uji Friedman berdasarkan *chi*-kuadrat.

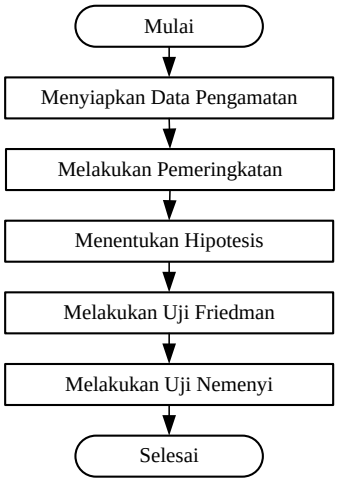
$$X_f^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \quad (1)$$

dengan r_i^j adalah peringkat ke-j dari k metode pada *dataset* ke-i dari N *dataset* dengan *degrees of freedom* (DF) (2).

$$DF = (k-1). \quad (2)$$

Keputusan hipotesisnya merupakan perbandingan antara nilai *chi*-kuadrat hitung (X_f^2) dan *chi-kuadrat* tabel ($X_{a(k-1)}^2$) yang dinotasikan sebagai berikut.

- Jika $X_f^2 < X_{a(k-1)}^2$, maka H_0 diterima dan H_a ditolak.
- Jika $X_f^2 > X_{a(k-1)}^2$, maka H_0 ditolak dan H_a diterima.



Gambar 2. Langkah-langkah perbandingan kinerja metode usulan.

Persamaan (3) digunakan pada uji Friedman berdasarkan *F-Distribution* (F_f).

$$F_f = \frac{(N-1)X_f^2}{N(k-1)-X_f^2}. \quad (3)$$

Persamaan (4) dan (5) digunakan untuk menghitung DF pada uji Friedman berdasarkan *F-Distribution* (F_f).

$$DF1 = (k-1) \quad (4)$$

$$DF2 = (k-1)(N-1). \quad (5)$$

Keputusan hipotesisnya merupakan perbandingan antara nilai *F-distribution* (F_f) dan *F-distribution* tabel ($F_{\alpha(k-1),(k-1)(N-1)}$), yang dinotasikan sebagai berikut.

- Jika $F_f > F_{\alpha(k-1),(k-1)(N-1)}$, maka H_0 ditolak dan H_a diterima.
- Jika $F_f < F_{\alpha(k-1),(k-1)(N-1)}$, maka H_0 diterima dan H_a ditolak.

Jika H_0 ditolak dan H_a diterima, proses dilanjutkan dengan uji Nemenyi untuk mengetahui metode yang secara berpasangan memiliki perbedaan siginifikan dalam penelitian ini. Perhitungan statistik uji Nemenyi dilakukan berdasarkan nilai *critical difference* (CD) [25] (6). Dua buah metode atau lebih dapat dikatakan memiliki perbedaan yang signifikan apabila nilai peringkat rata-ratanya lebih besar dari CD.

$$CD = q_\alpha\sqrt{\frac{K(K+1)}{6D}} \quad (6)$$

dengan q_α merupakan nilai kritis yang dipilih sesuai nilai taraf signifikansinya. Pedoman pemilihan nilai kritis ditunjukkan pada bagian selanjutnya. K adalah jumlah metode usulan yang dibandingkan, sedangkan D merupakan jumlah *dataset* yang digunakan dalam eksperimen setiap metode usulan.

C. EVALUASI HASIL EKSPERIMEN

Pada tahap evaluasi hasil eksperimen dilakukan evaluasi terhadap hasil eksperimen berupa perbandingan akurasi setiap metode usulan. Selanjutnya, ditarik kesimpulan dari penelitian yang dilakukan.

IV. HASIL DAN PEMBAHASAN

A. HASIL

Eksperimen yang dilakukan pada penelitian ini adalah mengoptimalkan algoritma KNN berdasarkan perbandingan

TABEL II
NILAI PARAMETER TETANGGA (*K*) DAN JUMLAH *OUTLIER* (*N*)

No.	Nama *Dataset*	Tetangga (*k*)	*Outlier* (*n*)
1.	Wisconsin Breast Cancer	5	30
2.	Glass	7	12
3.	Haberman	5	30
4.	Lymphography	5	10
5.	Parkinsons	3	15

analisis *outlier*. Eksperimen metode usulan KNN-*distance*, KNN-*density*, dan KNN-LOF dilakukan menggunakan RapidMiner. Hasil evaluasi ketiga metode usulan tersebut berupa akurasi menggunakan *confusion matrix* serta metode validasi 10-*fold cross validation*. Hasil rata-rata akurasi metode usulan setiap *dataset* dibandingkan tingkat signifikansinya menggunakan uji Friedman dan uji Nemenyi.

1) EKSPERIMEN KNN-*DISTANCE*

Operator deteksi *outlier* berbasis jarak pada RapidMiner mengidentifikasi *n outlier* dalam *dataset* berdasarkan jarak ke-*k* tetangga terdekatnya [28]. Operator ini melakukan pencarian *outlier* sesuai dengan pendekatan deteksi *outlier* yang direkomendasikan pada penelitian sebelumnya [29]. Penelitian tersebut mengusulkan formulasi *outlier* berbasis jarak yang didasarkan pada jarak suatu titik dari tetangga terdekatnya yang ke-*k*. Setiap titik diberi peringkat berdasarkan jaraknya ke tetangga terdekatnya yang ke-*k*. Lalu, *n* poin teratas dalam peringkat ini dinyatakan sebagai *outlier* [28]. Nilai *k* dan *n* dapat ditentukan berdasarkan jumlah tetangga dan jumlah parameter *outlier*.

Deteksi *outlier* (*distances*) dapat diatur berdasarkan jumlah parameter *k* tetangga dan jumlah *n outlier* dengan cara memilih parameter yang paling cocok [30], [31]. Pemilihan setiap parameter ditentukan menggunakan metode *trial and error*, yaitu mencoba satu per satu nilai parameter *k* tetangga bernilai ganjil (3, 5, dan 7). Dari nilai *k* tetangga tersebut, dicari nilai yang mengasilkan akurasi tertinggi [32]. Jumlah *outlier* yang akan dipilih harus disesuaikan dengan jumlah *instance* pada masing-masing *dataset*. Artinya setiap pencarian jumlah *outlier* tidak boleh lebih besar dari jumlah *instance*. Jumlah *k* tetangga dan *n outlier* masing-masing *dataset* dengan akurasi tinggi berdasarkan hasil *trial and error* dirangkum pada Tabel II.

Tabel II menunjukkan pemilihan nilai *k* tetangga dan jumlah *n outlier*. Sebagai contoh, KNN-*distance* akan menghasilkan kinerja terbaik pada Wisconsin Breast Cancer saat nilai *k*=5 dan *n*=30.

Proses penerapan model eksperimen KNN-*distance* dilakukan menggunakan beberapa operator, di antaranya *detect outlier* (*distances*), *filter examples*, *split data*, *multiply*, dan *cross validation*. Operator *filter examples* berfungsi untuk menghapus *outlier* yang telah terdeteksi. Operator ini bekerja dengan cara mengatur parameter filter yang telah ditambahkan pada kondisi *false*. Parameter filter yang diatur pada operator ini adalah *outlier*. Operator *split data* menghasilkan jumlah himpunan bagian yang diinginkan dari *dataset*. *Dataset* dibagi menjadi data latih (*training*) dan data uji (*testing*) dengan perbandingan 90% dan 10% [33]. Operator *cross validation* digunakan untuk memvalidasi kinerja model yang sudah diterapkan dengan 10-*fold* sebagai parameternya. Cara kerja 10-*fold cross validation* adalah membagi himpunan *dataset* menjadi 10-*fold* yang saling bebas: f_1, f_2, …., f_{10}, sehingga masing-masing *fold* berisi 1/10 bagian *dataset*. Selanjutnya,

TABEL III
HASIL RATA-RATA NILAI AKURASI KNN-*DISTANCE*

No.	Nama *Dataset*	KNN-*Distance*		
		K=3	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	96,08%	96,14%	95,99%
2.	Glass	82,68%	84,57%	84,79%
3.	Haberman	72,71%	73,94%	75,33%
4.	Lymphography	83,93%	83,17%	84,19%
5.	Parkinsons	90,95%	89,67%	88,17%

sepuluh himpunan *dataset*: D_1, D_2, …., D_{10} masing-masing berisi sembilan *fold* sebagai data latih dan satu *fold* sebagai data uji. Setiap *fold* akan menjadi data uji sebanyak satu kali [34]. Dalam proses *cross validation* terdapat dua halaman, yaitu halaman pelatihan dan halaman pengujian. Halaman pelatihan digunakan sebagai penerapan model algoritma KNN. Pemilihan nilai *k* ditentukan secara subjektif dan nilai *k* dianjurkan bernilai ganjil [32], [35]. Nilai *k* yang digunakan pada eksperimen ini adalah *k*=3, *k*=5, *k*=7 karena memiliki hasil akurasi yang tinggi dibandingkan dengan nilai *k* ganjil lainnya. Halaman pengujian berisi operator *apply model* dan *performances* (*classification*) yang berfungsi untuk mengetahui kinerja algoritma pada setiap *dataset*. Ukuran kinerja yang dipilih pada eksperimen KNN-*distance* adalah akurasi. Nilai rata-rata akurasi yang dihasilkan dari eksperimen KNN-*distance* menggunakan metode validasi kinerja model 10-*fold cross validation* ditunjukkan pada Tabel III.

Tabel III merupakan ringkasan rata-rata akurasi eksperimen KNN-*distance* yang dihasilkan berdasarkan 10-*fold cross validation* dengan nilai *k*=3, *k*=5, *k*=7. Dari semua *dataset*, Wisconsin Breast Cancer unggul dengan *k*=5 sebesar 96,14% dan Haberman menghasilkan rata-rata akurasi paling rendah dengan *k*=7, yakni sebesar 72,71%.

2) EKSPERIMEN KNN-*DENSITY*

Operator deteksi *outlier* berbasis kepadatan pada RapidMiner mengidentifikasi *outlier* dalam *dataset* berdasarkan kepadatan data. Semua objek yang memiliki setidaknya proporsi *p* dari semua objek yang lebih jauh dari jarak *D* dianggap sebagai *outlier* [28]. Proses penerapan model eksperimen KNN-*density* dilakukan menggunakan beberapa operator, di antaranya *detect outlier* (*densities*), *filter examples*, *split data*, *multiply*, dan *cross validation.*

Pada deteksi *outlier* (*densities*), pengaturan parameter jarak *D* dan proporsi *p* dilakukan dengan cara mencari parameter yang cocok [31]. Pemilihan setiap parameter ditentukan dengan metode *trial and error* nilai parameter jarak dengan nilai ganjil, yaitu 3, 5, 7, dan 9 [32], [35], serta proporsi *p* dengan nilai 0,1 sampai dengan 0,9. Nilai *k* yang digunakan pada eksperimen ini adalah *k*=3, *k*=5, *k*=7 karena memiliki hasil akurasi yang tinggi dibandingkan dengan nilai *k* ganjil lainnya. Berdasarkan *trial and error* nilai parameter jarak dan proporsi, didapatkan hasil nilai parameter yang dipilih dalam eksperimen, seperti ditampilkan pada Tabel IV. Nilai-nilai parameter jarak dan proporsi pada Tabel IV merupakan nilai parameter jarak dan proporsi yang menghasilkan akurasi tertinggi pada masing-masing *dataset*.

Penggunaan operator *filter examples*, *split data*, dan *cross validation* tidak berbeda dengan eksperimen penerapan KNN-*distance* yang sudah dijelaskan sebelumnya. Ukuran kinerja yang dipilih pada eksperimen KNN-*density* juga berupa akurasi. Nilai rata-rata akurasi yang dihasilkan dari eksperimen KNN-*density* menggunakan metode validasi kinerja model 10-*fold cross validation* ditunjukkan pada Tabel V.

TABEL IV
NILAI PARAMENTER JARAK DAN PROPORSI

No.	Nama *Dataset*	Jarak (*D*)	Proporsi (*p*)
1.	Wisconsin Breast Cancer	0,3	0,6
2.	Glass	0,3	0,7
3.	Haberman	0,3	0,8
4.	Lymphography	0,7	0,8
5.	Parkinsons	0,9	0,8

TABEL V
RATA-RATA NILAI AKURASI KNN-*DENSITY*

No.	Nama *Dataset*	KNN-*Density*		
		K=3	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	99,34%	99,34%	98,92%
2.	Glass	82,02%	85,14%	85,25%
3.	Haberman	70,79%	71,87%	73,08%
4.	Lymphography	83,13%	85,45%	82,18%
5.	Parkinsons	89,64%	89,80%	89,85%

Tabel V menyajikan hasil rata-rata akurasi KNN-*density* berdasarkan hasil 10-*fold cross validation* dengan nilai *k*=3, *k*=5 dan *k*=7 pada semua *dataset*. Wisconsin Breast Cancer menghasilkan rata-rata nilai akurasi tertinggi sebesar 99,34% saat bekerja dengan *k*=3 dan *k*=5, sedangkan nilai rata-rata akurasi terkecil, sebesar 70,79%, dihasilkan oleh Haberman dengan *k*=3.

3) *EKSPERIMEN KNN-LOF*

LOF didasarkan pada konsep kepadatan lokal. Lokalitas diberikan oleh *k* tetangga terdekat yang jaraknya digunakan untuk memperkirakan kerapatan. Dengan membandingkan kerapatan lokal suatu objek dengan kerapatan lokal tetangganya, daerah dengan kerapatan serupa dan titik yang memiliki kerapatan jauh lebih rendah dari tetangganya adalah *outlier* [28], [36].

Eksperimen KNN-LOF menggunakan beberapa operator, di antaranya *detect outlier* (LOF), *filter examples*, *split data*, *multiply*, dan *cross validation*. Operator *detect outlier* (LOF) pada RapidMiner memiliki parameter batas bawah dan batas atas titik minimal, masing-masing diatur pada nilai 10 dan 20 [37], yang diterapkan pada semua *dataset*. Penggunaan operator *filter examples*, *split data*, dan *cross validation* tidak berbeda dengan eksperimen penerapan KNN-*distance* dan KNN-*density* yang telah dibahas sebelumnya. Akurasi digunakan sebagai ukuran kinerja yang dipilih, sama seperti pada eksperimen KNN-*distance* dan KNN-*density*. Nilai rata-rata akurasi yang dihasilkan dari eksperimen KNN-LOF menggunakan metode validasi kinerja model 10-*fold cross validation* ditunjukkan pada pada Tabel VI.

Tabel VI menunjukkan rangkuman hasil rata-rata akurasi KNN-LOF yang dihasilkan berdasarkan 10-*fold cross validation* dengan nilai *k*=3, *k*=5, *k*=7. Dari semua *dataset*, Wisconsin Breast Cancer unggul dengan *k*=3 dan rata-rata akurasi sebesar 93,65%, sedangkan Haberman menghasilkan rata-rata akurasi paling rendah dengan *k*=3, yakni sebesar 67,50%.

D. PEMBAHASAN

Analisis *outlier* yang digunakan saat eksperimen berfungsi untuk menghapus *outlier* dalam *dataset*. Hasil dari *outlier* yang terdeteksi dan dihapus menggunakan analisis *outlier* (*distance*, *density*, dan LOF) ditunjukkan pada Tabel VII.

Tabel VII menunjukkan jumlah *outlier* yang terdeteksi pada masing-masing *dataset*. *Outlier* yang terdeteksi

TABEL VI
HASIL RATA-RATA NILAI AKURASI KNN-LOF

No.	Nama *Dataset*	KNN-LOF		
		K=3	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	93,65%	93,39%	93,41%
2.	Glass	81,27%	84,03%	82,02%
3.	Haberman	67,50%	69,76%	70,19%
4.	Lymphography	77,22%	74,74%	77,53%
5.	Parkinsons	89,54%	88,75%	87,78%

TABEL VII
OUTLIER YANG TERDETEKSI

No.	Nama *Dataset*	Jumlah *Instance*	*Outlier* Yang Terdeteksi			
			Density	*Distance*	*LOF*	
					Min	*Max*
1.	Wisconsin Breast Cancer	699	290	30	0,000	22,137
2.	Glass	214	0	12	0,926	2,617
3.	Haberman	306	7	30	0,931	3,301
3.	Lymphography	148	7	10	0,916	1,935
4.	Parkinsons	195	4	15	0,900	3,255

selanjutnya terhapus secara otomatis. KNN-*density* mendeteksi sebanyak 290 *outlier* pada Wisconsin Breast Cancer dan menjadi jumlah *outlier* terbanyak yang terdeteksi. Namun, KNN-*density* tidak menemukan *outlier* pada Glass, sehingga tidak ada penghapusan pada *dataset* tersebut. KNN-*distance* mendeteksi *outlier* terbanyak pada Wisconsin Breast Cancer serta Haberman, yakni 30 *outlier*, sedangkan Lymphography mendeteksi *outlier* dalam jumlah paling sedikit di antara *dataset* lainnya, yakni sepuluh *outlier*. Sementara itu, KNN-LOF hanya menampilkan nilai batas bawah dan batas atas. Nilai LOF tertinggi dianggap sebagai *outlier*. Dengan kata lain, nilai yang memiliki batas atas akan otomatis dianggap sebagai *outlier*. KNN-LOF mendeteksi *outlier* terbanyak pada Wisconsin Breast Cancer, dengan nilai batas atas 22,137, dan mendeteksi *outlier* paling sedikit pada Lymphography, dengan nilai batas atas 1,935.

1) PERBANDINGAN KINERJA METODE USULAN

Hasil evaluasi kinerja metode usulan menggunakan 10-*fold cross validation* dibandingkan menggunakan uji Friedman dan uji Nemenyi. Langkah-langkah perbandingan kinerja metode usulan adalah sebagai berikut.

Pertama, data pengamatan disiapkan. Data pengamatan yang digunakan adalah data hasil rata-rata akurasi setiap *dataset* dari masing-masing metode usulan yang ditampilkan pada Tabel III, Tabel V, dan Tabel VI. Selanjutnya, dilakukan pemeringkatan data pengamatan dengan cara mengurutkan nilai rata-rata akurasi setiap *dataset* dari masing-masing metode usulan. Kinerja metode usulan dengan akurasi terbaik pertama mendapatkan peringkat 1, metode usulan dengan akurasi terbaik kedua mendapatkan peringkat 2, dan seterusnya. Apabila ditemukan nilai akurasi yang sama, peringkat yang digunakan adalah peringkat rata-rata (RANK.AVG).

Tabel VIII menyajikan data pengamatan uji Friedman yang dihasilkan dari pemeringkatan nilai rata-rata akurasi setiap *dataset* dari masing-masing metode usulan. Sebagai contoh, Wisconsin Breast Cancer yang ditunjukkan pada Tabel V memiliki nilai akurasi tertinggi saat menggunakan metode KNN-*density*, dengan *k*=3 dan *k*=5, tetapi nilainya sama, yaitu 99,34%. Diasumsikan peringkatnya adalah peringkat 1 dan 2, sehingga peringkat rata-ratanya adalah 1,5. Sebagai contoh lain,

TABEL VIII
DATA PENGAMATAN UJI FRIEDMAN

No.	Nama *Dataset*	KNN-*Density*			KNN-*Distance*			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
1.	Wisconsin Breast Cancer	1,5	1,5	3	5	4	6	7	9	8
2.	Glass	7,5	2	1	6	4	3	9	5	7,5
3.	Haberman	6	5	3	4	2	1	9	8	7
4.	Lympo-graphy	5	1	6	3	4	2	8	9	7
5.	Parkinsons	5	3	2	1	4	8	6	7	9
Jumlah		25	12,5	15	19	18	20	39	38	38,5
Mean of rank		5	2,5	3	3,8	3,6	4	7,8	7,6	7,7
		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9

dataset Glass yang ditunjukkan pada Tabel V dan Tabel VI memiliki akurasi yang sama saat menggunakan metode KNN-*density* dengan *k*=3 dan KNN-LOF dengan *k*=7, yakni 82,02%. Diasumsikan peringkatnya adalah peringkat 7 dan 8, sehingga peringkat rata-ratanya adalah 7,5.

Langkah berikutnya adalah menentukan nilai taraf signifikansi (α). Pada penelitian ini, α yang digunakan adalah 0,05 (5%) dan 0,1 (10%). Penggunaan dua nilai taraf signifikansi dimaksudkan untuk melihat beberapa peluang hasil pengambilan keputusan hipotesis.

Selanjutnya, hipotesis ditentukan. Hipotesis dalam penelitian ini adalah sebagai berikut.

H_0: KNN-*density* = KNN-*distance* = KNN-LOF.

H_a: KNN-*density* ≠ KNN-*distance* ≠ KNN-LOF.

H_0 yang dimaksud adalah hipotesis yang menyatakan bahwa tidak ada perbedaan nilai rata-rata akurasi antara metode KNN-*distance*, KNN-*density*, dan KNN-LOF. Sementara itu, H_a adalah hipotesis yang menyatakan bahwa ada perbedaan nilai rata-rata akurasi antara metode KNN-*density*, KNN-*distance*, dan KNN-LOF.

Setelah menentukan hipotesis, langkah berikutnya adalah menghitung statistik uji Friedman. Uji Friedman berdasarkan *chi-kuadrat* dapat dihitung dengan diawali perhitungan nilai DF seperti pada (2) dan dilanjutkan menghitung nilai *chi-kuadrat* hitung seperti pada (1) dan *chi-kuadrat* tabel.

$$DF = (k-1) = (9-1) = 8$$

$$X_f^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

$$X_f^2 = \frac{12*5}{9(9+1)}\left[5^2 + 2{,}5^2 + 3^2 + 3{,}8^2 + 3{,}6^2 + 4^2 + 7{,}8^2 + 7{,}6^2 + 7{,}7^2 - \frac{9(9+1)^2}{4}\right]$$

$$X_f^2 = 0{,}66667[261{,}54 - 225]$$

$$X_f^2 = 24{,}36.$$

Chi-kuadrat tabel ($X^2_{\alpha(k-1)}$) dapat dicari dengan rumus Excel CHIINV(α;DF). Untuk nilai taraf signifikansi 5%, *chi-kuadrat* tabel yang dihasilkan adalah CHIINV(0,05;DF) = CHIINV(0,05;8) = 15,5073, sedangkan untuk nilai taraf signifikansi 10%, *chi-kuadrat* tabel yang dihasilkan adalah CHIINV(0,1;DF) = CHIINV(0,1;8) = 13,36157.

Dari hasil perhitungan yang dilakukan, diperoleh nilai *chi-kuadrat* hitung yang lebih besar daripada nilai *Chi-kuadrat* tabel, baik pada penggunaan taraf signifikansi 5% maupun 10%. Maka, keputusan hipotesisnya adalah $X_f^2 > X^2_{\alpha(k-1)}$ atau H_0 ditolak dan H_a diterima. Hal ini berarti terdapat perbedaan antara metode KNN-*distance*, KNN-*density*, dan KNN-LOF.

Uji Friedman berdasarkan *F-distribution* dapat dihitung dengan diawali perhitungan nilai $DF1$ seperti (4) dan $DF2$ seperti pada (5), dilanjutkan mencari nilai *F-distribution* (F_f) dan *F-distribution* tabel ($F_{\alpha(k-1),(k-1)(N-1)}$).

$$DF1 = (k-1) = (9-1) = 8$$

$$DF2 = (k-1)(N-1) = (9-1)(5-1) = 32$$

$$F_f = \frac{(N-1)X_f^2}{N(k-1) - X_f^2}$$

$$F_f = \frac{(5-1)*24{,}36}{5(9-1) - 24{,}36}$$

$$F_f = \frac{97{,}44}{15{,}64}$$

$$F_f = 6{,}23018.$$

Nilai *F-distribution* tabel dapat dihitung menggunakan rumus Microsoft Excel, yaitu FINV(α;($DF1$);($DF2$). Maka, *F-distribution* tabel untuk nilai taraf signifikansi 5% adalah $F_{a(k-1)(N-1)}$ = FINV(0,05;8;32) = 2,2444, sedangkan *F-distribution* tabel untuk nilai taraf signifikansi 10% adalah $F_{a(k-1)(N-1)}$ = FINV(0,1;8;32) = 1,8701.

Dari hasil perhitungan yang dilakukan, baik penggunaan taraf signifikansi 5% maupun 10% menghasilkan *F-distribution* lebih besar daripada *F-distribution* tabel. Maka, keputusan hipotesisnya adalah $F_f > F_{\alpha(k-1),(k-1)(N-1)}$ atau H_0 ditolak dan $H_a$diterima. Artinya terdapat perbedaan antara metode KNN-*distance*, KNN-*density*, dan KNN-LOF.

Berdasarkan statistik uji Friedman menggunakan *chi-kuadrat* dan *F-distribution*, didapatkan keputusan hipotesis menolak H_0 dan menerima H_a, sehingga disimpulkan bahwa ada perbedaan signifikan antara metode-metode usulan yang dibandingkan. Selanjutnya, dapat dilakukan uji Nemenyi untuk mengidentifikasi pasangan metode usulan yang paling berbeda secara signifikan. Tahapan uji Nemenyi yang dilakukan dalam penelitian ini adalah sebagai berikut.

Langkah pertama adalah menyiapkan data pengamatan yang sudah disajikan pada Tabel VIII. Data pengamatan pada tabel tersebut merupakan data pengamatan yang sama pada uji Friedman. Langkah berikutnya adalah menghitung nilai CD seperti pada (6) berdasarkan nilai kritis (q_a) yang dipilih. Nilai kritis yang dapat digunakan pada uji Nemenyi disajikan pada Tabel IX.

Tabel IX menunjukkan penggunakan nilai kritis yang dapat dipakai dalam uji Nemenyi. Dalam penelitian ini, nilai kritis yang dipakai adalah taraf signifikansi (α) 0,05 dan taraf signifikansi (α) 0,1 pada *classifier* 9, yakni 3,102 dan 2,855. *Classifier* 9 dipilih karena berdasarakan Tabel VII, terdapat sembilan nilai akurasi yang dibandingkan. Nilai CD sesuai taraf signifikansinya yang diperoleh adalah sebagai berikut.

$$CD = q_{0,05}\sqrt{\frac{K(K+1)}{6D}} = 3{,}102\sqrt{\frac{9(9+1)}{6*5}} = 5{,}3728$$

$$CD = q_{0,1}\sqrt{\frac{K(K+1)}{6D}} = 2{,}855\sqrt{\frac{9(9+1)}{6*5}} = 4{,}9450.$$

TABEL IX
NILAI KRITIS UNTUK UJI NEMENYI

Classifier	$q_{0,05}$ (5%)	$q_{0,1}$ (10%)
2	1,960	1,645
3	2,343	2,052
4	2,569	2,291
5	2,728	2,459
6	2,850	2,589
7	2,949	2,693
8	3,031	2,780
9	3,102	2,855
10	3,164	2,920

TABEL X
DATA PENGAMATAN UJI NEMENYI

		KNN-*Density*			KNN-*Distance*			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
KNN-*density*	*K*=3	0	2,5	2	1,2	1,4	1	2,8	2,6	2,7
	K=5	2,5	0	0,5	1,3	1,1	1,5	5,3	5,1	5,2
	K=7	2	0,5	0	0,8	0,6	1	4,8	4,6	4,7
KNN-*distance*	*K*=3	1,2	1,3	0,8	0	0,2	0,2	4	3,8	3,9
	K=5	1,4	1,1	0,6	0,2	0	0,4	4,2	4	4,1
	K=7	1	1,5	1	0,2	0,4	0	3,8	3,6	3,7
KNN-LOF	*K*=3	2,8	5,3	4,8	4	4,2	3,8	0	0,2	0,1
	K=5	2,6	5,1	4,6	3,8	4	3,6	0,2	0	0,1
	K=7	2,7	5,2	4,7	3,9	4,1	3,7	0,1	0,1	0
Critical Difference ($q_{0,05}$) = 5,3728										
Critical Difference ($q_{0,1}$) = 4,9450										

Langkah berikutnya adalah menghitung selisih *mean of rank* antara dua metode yang dibandingkan. Dengan kata lain, uji Nemenyi akan menampilkan perbandingan berpasangan dari metode usulan. Sebuah metode dikatakan berbeda signifikan jika selisih *mean of rank* kedua metode yang dibandingkan lebih besar dari nilai CD yang dihasilkan. Pada Tabel VIII ditunjukkan *mean of rank* (R_1 sampai dengan R_9). Sebagai contoh, selisih *mean of rank* antara KNN-*density* (*k*=3) terhadap KNN-*density* (*k*=5) adalah 2,5, yang diperoleh dari selisih *mean of rank* R_1, yaitu 5, dengan *mean of rank* R_2, yaitu 2,5. Data keseluruhan pengamatan uji Nemenyi berdasarkan selisih *mean of rank* ditunjukkan pada Tabel X.

Tabel X menunjukkan perbandingan antara metode-metode yang dihasilkan dari selisih *mean of rank* setiap metode terhadap metode yang lain. Sebagai contoh, perbandingan antara KNN-*density* *k*=3 dengan KNN-*density* *k*=3 hasilnya adalah 0, KNN-*density* *k*=3 dengan KNN-*density* *k*=5 hasilnya adalah 2,5, KNN-*density* *k*=3 dengan KNN-*density* *k*=7 hasilnya adalah 2, dan seterusnya.

Langkah terakhir dari uji Nemenyi adalah membandingkan hasil data pengamatan pada Tabel X dengan CD yang sudah dihasilkan pada tahap sebelumnya untuk melihat signifikansi perbedaan antara metode-metode yang diusulkan. Hasil perbandingan antara rata-rata peringkat dari data pengamatan dengan hasil CD memiliki perbedaan, yaitu bernilai *No* dan *Yes*. Jika nilai CD lebih besar dari nilai rata-rata peringkat, nilainya menjadi *No*, yang artinya tidak ada perbedaan secara signifikan antara metode-metode yang diusulkan (KNN-*density* = KNN-*distance* = KNN-LOF). Jika nilai CD lebih kecil dari nilai rata-rata peringkat, nilainya *Yes*, yang artinya metode usulan berbeda secara signifikan (KNN-*density* ≠ KNN-*distance* ≠ KNN-LOF). Hasil perbandingan uji Nemenyi ditunjukkan pada Tabel XI untuk taraf signifikansi 5% dan Tabel XII untuk taraf signifikansi 10%.

TABEL XI
HASIL UJI NEMENYI ($\alpha = 0,05$)

		KNN-*Density*			KNN-*Distance*			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
KNN-*density*	*K*=3	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-*distance*	*K*=3	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-LOF	*K*=3	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*

TABEL XII
HASIL UJI NEMENYI ($\alpha = 0,1$)

		KNN-*Density*			KNN-*Distance*			KNN-LOF		
		K=3	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*	*K=3*	*K=5*	*K=7*
KNN-*density*	*K*=3	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-*distance*	*K*=3	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
KNN-LOF	*K*=3	*No*	*Yes*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=5	*No*	*Yes*	*No*	*No*	*No*	*No*	*No*	*No*	*No*
	K=7	*No*	*Yes*	*No*	*No*	*No*	*No*	*No*	*No*	*No*

Berdasarkan Tabel XI, dapat disimpulkan bahwa hasil uji Nemenyi dengan taraf signifikansi 5% menunjukkan tidak adanya perbedaan yang signifikan antarmetode usulan. Namun, berdasarkan Tabel XII, dapat disimpulkan bahwa hasil uji Nemenyi dengan taraf signifikansi 10% adalah KNN-*density* *k*=5 berbeda secara signifikan dengan KNN-LOF *k*=3, *k*=5, dan *k*=7, serta KNN-LOF *k*=3, *k*=5, dan *k*=7 berbeda signifikan terhadap KNN-*density* *k*=5. Dengan demikian, terbukti bahwa hasil uji Nemenyi menggunakan taraf signifikansi 10% menunjukkan adanya perbedaan metode usulan secara signifikan.

V. KESIMPULAN

Hasil penelitian menunjukkan bahwa KNN-*density* menjadi metode yang menghasilkan rata-rata akurasi tinggi sebanyak tiga kali: pada Wisconsin Breast Cancer dengan nilai rata-rata akurasi sebesar 99,34% pada *k*=3 dan *k*=5; pada Glass dengan rata-rata akurasi sebesar 85,25% pada *k*=7; dan pada Lymphography dengan rata-rata akurasi sebesar 85,45% pada *k*=5. Berdasarkan uji Friedman menggunakan taraf signifikansi 5% dan 10%, diperoleh hasil bahwa hipotesis H_0 ditolak dan H_a diterima. Ini berarti terdapat perbedaan antara KNN-*density*, KNN-*distance*, dan KNN-LOF. Selanjutnya, berdasarkan uji Nemenyi dengan taraf signifikansi 5%, tidak terdapat perbedaan yang signifikan antarmetode usulan. Namun, saat digunakan taraf signifikansi 10%, terbukti bahwa ada perbedaan yang signifikan antara KNN-*density* dengan KNN-LOF. Dari hasil rata-rata akurasi, dapat disimpulkan bahwa metode usulan KNN-*density* mampu mengoptimasi algoritma KNN dengan mendeteksi dan menghapus *outlier* menggunakan analisis *outlier* berbasis kepadatan (*density*). Hal tersebut merupakan kontribusi yang dapat digunakan sebagai jawaban pada masalah penelitian dan tujuan penelitian pada penelitian ini.

KONFLIK KEPENTINGAN

Penulis menyatakan bahwa selama melaksanakan penelitian dan penulisan artikel ilmiah dengan judul "Optimasi Algoritma *K-Nearest Neighbors* Berdasarkan Perbandingan Analisis *Outlier* (Berbasis Jarak, Kepadatan, LOF)", tim penulis tidak memiliki konflik kepentingan dengan pihak mana pun.

KONTRIBUSI PENULIS

Konseptualisasi, Fitri Ayuning Tyas dan Mahda Nurayuni; metodologi, Mahda Nurayuni; perangkat lunak, Mahda Nurayuni; validasi, Fitri Ayuning Tyas, Mahda Nurayuni, dan Hidayatur Rakhmawati; penulisan—penyusunan draf asli, Fitri Ayuning Tyas, Mahda Nurayuni, Hidayatur Rakhmawati; penulisan—peninjauan dan penyuntingan, Fitri Ayuning Tyas; visualisasi, Fitri Ayuning Tyas; pengawasan, Fitri Ayuning Tyas; pendanaan, Fitri Ayuning Tyas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada rekan-rekan di Laboratorium Komputer STMIK Muhammadiyah Paguyangan Brebes, sebagai bagian dari Program Studi Sistem Informasi yang telah berpartisipasi dalam penyelesaian penelitian ini.

REFERENSI

[1] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Massachusetts, AS: Morgan Kaufmann, 2012.

[2] O. Alghushairy, R. Alsini, T. Soule, dan X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data Cogn. Comput.*, vol. 5, no. 1, hal. 1–24, Mar. 2021, doi: 10.3390/bdcc5010001.

[3] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. Heidelberg, Jerman: Springer, 2011.

[4] I.H. Witten, E. Frank, dan M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Massachusetts, USA: Morgan Kaufmann, 2011.

[5] C.C. Aggarwal, *Data Mining.* New York, ~~NY,~~ USA: Springer, 2015.

[6] H. Liu dan S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining," *J. Syst. Softw.*, vol. 85, no. 5, hal. 1067–1024, Mei 2012, doi: 10.1016/j.jss.2011.12.019.

[7] D. Armiady, "Analisis Metode DBSCAN (density-based spatial clustering of application with noise) dalam mendeteksi data outlier," *JURIKOM (J. Ris. Komputer)*, vol. 9, no. 6, hal. 2158–2164, Des. 2022, doi: 10.30865/jurikom.v9i6.5080.

[8] R. Silvi, "Analisis cluster dengan data outlier menggunakan centroid linkage dan k-means clustering untuk pengelompokan indikator HIV/AIDS di Indonesia," *J. Mat. MANTIK*, vol. 4, no. 1, pp. 22–31, May 2018, doi: 10.15642/mantik.2018.4.1.22-31.

[9] M.Y. Pusadan, "Outlier detection pada set data flight recording (pre-processing sumber data ADS-B)," *Seminar Nas. Teknol. Inf. Multimedia 2015*, 2015, hal. 2.1-31–2.1-36.

[10] J. Abellán, J.G. Castellano, dan C.J. Mantas, "A new robust classifier on noise domains: Bagging of credal C4.5 trees," *Complexity, vol. 2017, pp. 1–17, Dec. 2017, Art. no. 9023970, doi: 10.1155/2017/9023970*.

[11] A. Duraj dan P.S. Szczepaniak, "Outlier detection in data streams — A comparative study of selected methods," *Procedia Comput. Sci.*, vol. 192, hal. 2769–2778, Okt. 2021, doi: 10.1016/j.procs.2021.09.047.

[12] S. Sugidamayatno dan D. Lelono, "Outlier detection credit card transactions using local outlier factor algorithm (LOF)," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 13, no. 4, hal. 409–420, Okt. 2019, doi: 10.22146/ijccs.46561.

[13] X. Xu, H. Liu, L. Li, dan M. Yao, "A comparison of outlier detection techniques for high-dimensional data," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, hal. 652–662, Jan. 2018, doi: 10.2991/ijcis.11.1.50.

[14] T. Sangeetha dan G. Mary A, "A fuzzy proximity relation approach for outlier detection in the mixed dataset by using rough entropy-based weighted density method," *Soft Comput. Lett.*, vol. 3, hal. 1–12, Des. 2021, doi: 10.1016/j.socl.2021.100027.

[15] H. Xu, L. Zhang, P. Li, and F. Zhu, "Outlier detection algorithm based on k-nearest neighbors-local outlier factor," *J. Algorithms Comput. Technol.*, vol. 16, hal. 1–12, Mar. 2022, doi: 10.1177/17483026221078111.

[16] X. Wu dkk., "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, hal. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[17] Z. Deng dkk., "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, hal. 143–148, Jun. 2016, doi: 10.1016/j.neucom.2015.08.112.

[18] S. Zhang dkk., "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, hal. 1774–1785, Mei 2018, doi: 10.1109/TNNLS.2017.2673241.

[19] J. Ning, L. Chen, C. Zhou, dan Y. Wen, "Parameter k search strategy in outlier detection," *Pattern Recognit. Lett.*, vol. 112, hal. 56–62, Sep. 2018, doi: 10.1016/j.patrec.2018.06.007.

[20] O. Maimon dan L. Rokach, *Data Mining and Knowledge Discovery Handbook.* New York, USA: Springer, 2010.

[21] P.A. Ariawan, "Optimasi pengelompokan data pada ketode k-means dengan analisis outlier," *J. Nas. Teknol. Sist. Inf.*, vol. 5, no. 2, hal. 88–95, Agu. 2019, doi: 10.25077/TEKNOSI.v5i2.2019.88-95.

[22] H.C. Mandhare dan S.R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," *2017 Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, 2017, hal. 931–935, doi: 10.1109/ICCONS.2017.8250601.

[23] J. Yang, S. Rahardja, dan P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognit.*, vol. 115, hal. 1–11, Jul. 2021, doi: 10.1016/j.patcog.2021.107874.

[24] H. Wang, M.J. Bah, dan M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, hal. 107964–108000, Agu. 2019, doi: 10.1109/ACCESS.2019.2932769.

[25] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, hal. 1–30, Jan. 2006.

[26] I.C. Negara dan A. Prabowo, "Penggunaan uji chi–square untuk mengetahui pengaruh tingkat pendidikan dan umur terhadap pengetahuan penasun mengenai HIV–AIDS di Provinsi DKI Jakarta," *Pros. Senamantra (Seminar Nas. Mat. Terapannya)*, 2018, pp. 1–8.

[27] L.F. Obe, D. Lalang, V. Lakapeni, dan D. Fatin, "Pengaruh jumlah anak terhadap pendapatan hasil perkebunan kemiri di Desa Maikang Kecamatan Alor Selatan tahun 2020 menggunakan metode chi kuadrat," *J. Ilm. Wahana Pendidikan*, vol. 7, no. 6, hal. 378–384, Okt. 2021, doi: 10.5281/zenodo.5644452.

[28] F. Akthar dan C. Hahne, *RapidMiner 5 Operator Reference*. 2012.

[29] S. Ramaswamy, R. Rastogi, dan K. Shim, "Efficient algorithms for mining outliers from large data sets," *Proc. 2000 ACM SIGMOD Int. Conf. Manag. Data*, 2000, hal. 427–438, doi: 10.1145/342009.335437.

[30] Z.A Bakar, R. Mohemad, A. Ahmad, dan M.M. Deris, "A comparative study for outlier detection techniques in data mining," *2006 IEEE Conf. Cybern. Intell. Syst.*, 2006, hal. 1–6, doi: 10.1109/ICCIS.2006.252287.

[31] B. Tang dan H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, hal. 171–180, Jun. 2017, doi: 10.1016/j.neucom.2017.02.039.

[32] D. Kartini dkk., "Perbandingan nilai k pada klasifikasi pneumonia anak balita," *J. Komputasi*, vol. 10, no. 1, hal. 47–53, Apr. 2022, doi: 10.23960%2Fkomputasi.v10i1.2965.

[33] R.M. Candra dan A.N Rozana, "Klasifikasi komentar bullying pada Instagram menggunakan metode k-nearest neighbor," *IT J. Res. Dev.*, vol. 5, no. 1, hal. 45–52, Jul. 2020, doi: 10.25299/itjrd.2020.vol5(1).4962.

[34] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Massachusetts, USA: Morgan Kaufmann, 2012.

[35] M. Rivki dan A.M. Bachtiar, "Implementasi algoritma k-nearest neighbor dalam pengklasifikasian follower Twitter yang menggunakan bahasa Indonesia," *J. Sist. Inf. (J. Inf. Syst.)*, vol. 13, no. 1, hal. 31–37, Apr. 2017, doi: 10.21609/jsi.v13i1.500.

[36] A. Mahendra, "Pentapisan dan deteksi data outlier dalam proses sistem akusisi data pada proses sintering," *Arsitron*, vol. 6, no. 1, hal. 1–7, Jun. 2015.

[37] D. Handriyadi, M.A. Bijaksana, dan E.B. Setiawan, "Analisis perbandingan clustering-based, distance-based dan density-based dalam mendeteksi outlier," *Seminar Nas. Apl. Teknol. Inf. (SNATI)*, 2009, hal. 101–108.