Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 6, Desember 2022, hlm. 1709-1716

DOI: https://doi.org/10.20884/1.jutif.2022.3.6.471
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON OF FEATURE SELECTION TO PERFORMANCE IMPROVEMENT OF K-NEAREST NEIGHBOR ALGORITHM IN DATA CLASSIFICATION

Iswanto[1], Tulus[*2], Poltak Sihombing[3]

[1,2,3]Program Studi S2 Teknik Informatika, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Indonesia
Email: [1]iswan1512@gmail.com, [2]tulus@usu.ac.id, [3]poltak@usu.ac.id

(Naskah masuk: 13 Juli 2022, Revisi : 17 Juli 2022, diterbitkan: 26 Desember 2022)

Abstract

One of the most widely used data classification methods is the K-Nearest Neighbor (K-NN) algorithm. Classification of data in this method is carried out based on the calculation of the closest distance to the training data as much as the value of K from its neighbors. Then the new data class is determined using the most votes system from the number of K nearest neighbors. However, the performance of this method is still lower than other data classification methods. The cause is the use of the most voting system in determining new data classes and the influence of features less relevant to the dataset. This study compares several feature selection methods in the data set to see their effects on the performance of the K-NN algorithm in data classification. The feature selection methods in this research are Information gain, Gain ratio, and Gini index. The method was tested on the Water Quality dataset from the Kaggle Repository to see the most optimal feature selection method. The test results on the dataset show that the use of the feature selection method affects to increase the performance of the K-NN algorithm. The average increase in the accuracy value obtained from the value of K=1 to K=15 is the Information Gain increased by 1.17%, Gain ratio increased by 0.69%, and the Gini index increased by 1.19%. The highest accuracy value in the classification of the Water Quality dataset is 89.66% at K=13 with the Information Gain feature selection method.

Keywords: *Gain Ratio, Gini Index, Information Gain, K-Nearest Neighbor.*

PERBANDINGAN SELEKSI FITUR TERHADAP PENINGKATAN KINERJA ALGORITMA K-NEAREST NEIGHBOR DALAM KLASIFIKASI DATA

Abstrak

Salah satu metode klasifikasi data yang banyak digunakan adalah algoritma K-Nearest Neighbor (K-NN). Klasifikasi data pada metode ini dilakukan berdasarkan perhitungan jarak terdekat dengan data latih sebanyak nilai K dari tetangganya, kemudian kelas data baru ditentukan menggunakan sistem suara terbanyak dari jumlah K tetangga terdekat yang ditentukan. Namun kinerja dari metode ini masih lebih rendah jika dibandingkan dengan metode klasifikasi data lainnya. Hal ini disebabkan oleh penggunaan sistem *vote majority* dalam penentuan kelas data baru dan pengaruh dari fitur-fitur yang kurang relevan pada dataset. Penelitian ini membandingkan beberapa metode seleksi fitur pada dataset untuk melihat pengaruhnya terhadap kinerja algoritma K-NN dalam klasifikasi data. Metode seleksi fitur yang digunakan adalah *Informatian Gain, Gain Ratio, dan Gini Index*. Metode tersebut diujikan pada dataset *Water Quality* yang diambil dari *Kaggle Repository* untuk melihat metode seleksi fitur yang paling optimal. Dari hasil pengujian pada dataset menunjukkan bahwa penggunaan metode seleksi fitur memberikan pengaruh terhadap peningkatan kinerja algoritma K-NN. Rata-rata peningkatan nilai akurasi yang diperoleh dari nilai K=1 sampai K=15 yaitu *Information Gain* meningkat 1,17%, *Gain ratio* meningkat 0,69% dan *Gini Index* meningkat 1,19%. Nilai akurasi tertinggi pada klasifikasi dataset *Water Quality* diperoleh sebesar 89,66% pada nilai K=13 dengan metode seleksi fitur *Information Gain.*

Kata kunci: *Gain Ratio, Gini Index, Information Gain, K-Nearest Neighbor.*

1. PENDAHULUAN

Pengolahan data merupakan hal yang sangat penting dalam kehidupan sehari-hari. Salah satu teknik *Data Mining* yang banyak digunakan adalah *Classification*. Sekumpulan data pada umumnya menggunakan metode klasifikasi untuk menggolongkan data pada kelas-kelas yang ditentukan[1]. Salah satu metode klasifikasi data yang

banyak digunakan yaitu algoritma K-Nearest Neighbor (KNN). Metode ini mengklasifikasikan data berdasarkan jarak terdekat dari data uji ke data latih dengan banyak nilai K dari tetangganya [2]. Sedangkan untuk menentukan kelas data baru, algoritma K-NN menggunakan sistem suara terbanyak (*vote majority*) dari jumlah K tetangga terdekat yang ditentukan [3]. Tetapi dalam penerapannya pada proses klasifikasi data metode ini memiliki masalah terhadap kinerja yang cenderung lebih rendah bila dibandingkan dengan metode klasifikasi data lainnya yang ditunjukkan dari nilai akurasi yang dihasilkan. Hal ini dapat dilihat dari hasil perbandingan kinerja metode klasifikasi data antara algoritma *Naïve Bayes*, K-NN dan *Support Vector Machine(SVM)* pada penelitian sebelumnya yang menunjukkan bahwa *Naïve Bayes* memiliki kinerja yang lebih unggul jika dibandingkan dengan K-NN dan SVM. Hasil klasifikasianya menunjukkan bahwa Naïve Bayes memiliki akurasi sebesar 31,5%, diikuti oleh SVM sebesar 28,5% dan terakhir K-NN sebesar 23,8%. Berdasarkan penelitian tersebut Algoritma K-NN memiliki kinerja paling rendah [4].

Rendahnya nilai akurasi algoritma K-NN disebabkan pada penentuan kelas data baru yang menggunakan sistem vote majority sehingga menjadi tidak rasional ketika jarak setiap tetangga terdekat sangat berbeda terhadap jarak data uji [5]. Selain itu rendahnya akurasi juga disebabkan oleh penggunaan fitur pada dataset yang kurang relevan sehingga perlu dilakukan tahapan *preprocessing* data untuk memilih fitur-fitur yang sesuai [6]. Proses seleksi fitur memberikan pengaruh terhadap kinerja klasifikasi. Proses ini sangat penting dalam mengenali pola dan data analisis yang bertujuan untuk memilih fitur terbaik dari seluruh fitur yang tersedia. Dengan demikian dapat mengurangi dimensi data yang tinggi dan mampu memberikan solusi dari "*curse of dimensionality*" yang mampu meningkatkan kinerja algoritma K-NN [7].

Upaya meningkatkan kinerja algoritma K-NN menggunakan metode seleksi fitur juga telah dilakukan oleh peneliti sebelumnya dengan menggunakan metode seleksi fitur Gain Ratio yang diterapkan pada dataset yang bersumber dari *UCI Machine Learning repository*. Hasil penelitian tersebut menunjukkan bahwa penggunaan metode seleksi fitur Gain Ratio terbukti mampu meningkatkan rata-rata nilai akurasi pada klasifikasi data sebesar 4,09% dengan rata-rata nilai akurasi sebesar 90,58% untuk seluruh nilai K yang ditentukan [8].

Selain itu ada juga penelitian yang membandingkan beberapa algoritma klasifikasi (*Deep Learning, Decision Tree, K-NN, Naïve Bayes*) dengan menerapkan seleksi fitur *Information Gain* dan *Chi Square* pada analisis sentiment *Bitcoin*. Hasil yang diperoleh dari penelitian ini menunjukkan bahwa penggunaan seleksi fitur *Information Gain* memperoleh hasil akurasi paling baik dengan rata-rata nilai akurasi sebesar 63,79% jika dibandingkan dengan *Chi Square* yang memperoleh rata-rata akurasi 63,48%. Sedangkan untuk nilai akurasi tertinggi diperoleh pada algoritma *Deep Learning* dengan seleksi fitur *Information Gain* yaitu sebesar 78,63%. Hal ini menunjukkan bahwa penggunaan seleksi fitur pada penelitian tersebut juga mampu meningkatkan kinerja klasifikasi data [9].

Penelitian lain dalam upaya meningkatkan kinerja alogoritma K-NN adalah dengan menerapkan metode seleksi fitur *Gini Index* pada klasifikasi data kemajuan hasil belajar siswa. Pada penelitian tersebut diperoleh hasil bahwa penggunaan metode seleksi fitur *Gini Ind*ex mampu meningkatkan kinerja algoritma K-NN sebesar 2,45% dengan rata-rata nilai akurasi sebesar 76,52% sedangkan pada K-NN tanpa seleksi fitur hanya diperoleh rata-rata nilai akurasi sebesar 74,07% [10].

Metode seleksi fitur lain yang diterapkan oleh peneliti sebelumnya pada algoritma K-NN adalah *Symmetrical* Uncertainty. Dari hasil penelitian diperoleh bahwa penggunaan seleksi fitur *Symmetrical Uncertainty* pada K-NN mampu meningkatkan kinerja K-NN rata-rata sebesar 3,00%. Rata-rata nilai akurasi pada K-NN sebesar 66,35% dan K-NN dengan seleksi fitur *Symmetrical Uncertainty* sebesar 69,35% [11].

Selanjutnya ada juga peneliti yang menggunakan metode seleksi fitur untuk meningkatkan nilai akurasi algoritma K-NN. Peneliti tersebut membandingkan dua metode seleksi fitur yaitu *Gain Ratio* dan *Principal Component Analysis* (PCA). Berdasarkan hasil penelitiannya disebutkan bahwa seleksi fitur dapat meningkatkan akurasi K-NN dengan rata-rata peningkatan sebesar 19,15% dan metode seleksi fitur yang lebih baik dalam meningkatkan akurasi adalah *Gain Ratio*[12].

Berdasarkan beberapa penelitian sebelumnya menunjukkan bahwa penggunaan metode seleksi fitur yang diusulkan oleh peniliti mampu meningkatkan kinerja algoritma K-NN yang dapat dilihat dari nilai akurasi yang meningkat pula. Pada penelitian ini, penulis ingin membandingkan beberapa metode seleksi fitur yang telah digunakan oleh peneliti sebelumnya pada algoritma K-NN. Selanjutnya dilakukan analisa terhadap kinerja yang dihasilkan berdasarkan pengujian menggunakan dataset.

Metode seleksi fitur yang digunakan pada penelitian ini adalah *Information Gain, Gain Ratio* dan *Gini Index.* Sedangkan dataset yang digunakan adalah *Water Quality* yang bersumber dari *Kaggle Repository.* Dataset ini memiliki record sebanyak 7.996 dengan fitur berjumlah 20. Fitur tersebut nantinya akan diseleksi menggunakan ketiga metode seleksi fitur untuk membuang atau mereduksi fitur-fitur yang kurang relevan sebelum dilakukan proses klasifikasi data menggunakan algoritma K-NN. Berkurangnya fitur pada dataset tentunya juga akan mempengaruhi kinerja algoritma K-NN dari segi

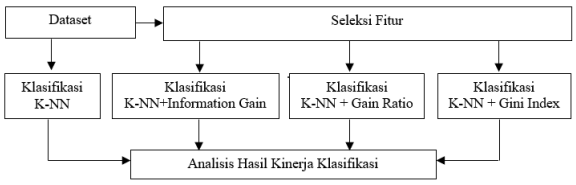
waktu komputasi yang semakin cepat dalam proses klasifikasi data.

Dengan adanya penelitian ini diharapkan dapat diketahui metode seleksi fitur mana yang paling optimal dari ketiga metode seleksi fitur yang digunakan untuk meningkatkan kinerja algoritma K-NN yang dapat dilihat dari hasil pengujiannya. Sehingga dapat membarikan kontribusi pada ilmu pengetahuan dalam bidang klasifikasi data khususnya pada pengembangan metode algoritma K-NN dengan memberikan solusi yang lebih optimal.

2. METODE PENELITIAN

2.1. Contoh Sub-Bab Pertama

Dalam membandingkan beberapa metode seleksi fitur pada algoritma K-NN untuk menganalisa kinerja yang dihasilkan, penulis melakukan tahapan-tahapan penelitian yang ditunjukkan pada gambar di bawah ini :



Gambar 1. Tahapan penelitian yang dilakukan

Berdasarkan gambar 1 di atas dapat dijelaskan tahapan-tahapan yang dilakukan dalam penelitian ini adalah sebagai berikut :

a. Memilih dataset yang sesuai untuk diujikan pada metode seleksi fitur dan metode klasifikasi data.
b. Melakukan proses seleksi fitur berdasarkan perhitungan bobot dari setiap fiturnya menggunakan metode seleksi fitur *information gain, gain ratio* dan *gini index*.
c. Pengujian klasifikasi data dengan 4 tahapan proses klasifikasi yaitu menggunakan algoritma K-NN tanpa seleksi fitur, K-NN dengan seleksi fitur *Information Gain*, K-NN dengan seleksi fitur *Gain Ratio* dan K-NN dengan seleksi fitur *Gini Index*.
d. Pengukuran kinerja dari hasil pengujian 4 proses klasifikasi dengan melihat nilai akurasi. Untuk mengitung tingkat akurasi dapat menggunakan persamaan berikut [13] :

$$Akurasi = \frac{Jumlah\ prediksi\ Benar}{Jumlah\ seluruh\ data}\ x\ 100\% \qquad (1)$$

2.2. Dataset yang Digunakan

Dataset yang digunakan pada penelitian ini adalah dataset *Water Quality* yang bersumber dari *Kaggle Repository*. Dataset ini memiliki 20 fitur dengan baris data berjumlah 7.996 record yang diklasifikasikan ke dalam 2 kelas data yaitu aman dan tidak aman untuk digunakan. Dalam proses klasifikasi dataset dibagi secara langsung menjadi 2 bagian dengan ketentuan 70% data dijadikan sebagai data latih dan 30% data dijadikan sebagai data uji.

Tabel 1. Rincian dataset yang digunakan

Dataset	Tipe	Fitur	Record	Kelas	Sumber
Water Quality	Real	20	7.996	2	*Kaggle Repository*

2.3. Seleksi Fitur

Seleksi fitur bertujuan untuk mengurangi atau mengeliminasi fitur yang kurang relevan. Dengan pengurangan fitur tersebut dalam proses klasifikasi data tentunya juga akan mengurangi waktu komputasi dalam proses klasifikasi data sehingga algoritma akan menjadi lebih cepat dalam menangani data [14]. Pada penelitian ini menggunakan tiga metode seleksi fitur yaitu *Information Gain, Gain Ratio dan Gini Index.*

Pada proses seleksi fitur dilakukan perhitungan nilai bobot dari setiap fitur pada dataset menggunakan metode seleksi fitur yang telah ditentukan. Selanjutnya fitur diurutkan dari bobot tertinggi sampai bobot terendah. Berdasarkan nilai bobot fitur kemudian dihutung proporsi masing-masing fitur dengan persamaan berikut:

$$Proporsi = \frac{Bobot\ fitur}{Total\ Bobot\ seluruh\ fitur}\ x\ 100\% \qquad (2)$$

Pada penelitian ini untuk menentukan fitur yang dipilih dalam proses klasifikasi data ditentukan berdasarkan total proporsi fitur maksimal yang ditetapkan sebesar 90% [12]. Fitur dipilih mulai dari proporsi yang tertinggi hingga mencapai batas total maksimal yang ditentukan. Jika fitur-fitur yang terpilih telah mencapai total proporsi 90% maka fitur lainnya akan dibuang.

2.3.1. Information Gain

Information Gain adalah salah satu metode seleksi fitur yang sederhana yaitu dengan merangking fitur. Metode ini banyak digunakan pada pengkategorian teks, data citra dan data *microarray* [15]. *Information Gain* mampu mengurangi *noise* yang ditimbulkan oleh fitur-fitur yang kurang relevan dan juga mampu mendeteksi fitur-fitur yang memiliki informasi terbanyak sesuai dengan kelas tertentu. Pemilihan fitur dilakukan dengan menghitung nilai *entropy* yang merupakan ukuran ketidakpastian kelas dengan menggunakan probabilitas kejadian atau fitur tertentu [16]. Berikut ini adalah langkah dalam menentukan nilai *Information Gain* pada setiap fitur:

a. Menghitung nilai *Entropy* setiap fitur dengan rumus persamaan :

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * log_2 pi \qquad (3)$$

Dimana *S* adalah Himpunan kasus, *n* banyaknya partisi *S*, dan *pi* proporsi dari *Si* terhadap *S.*

b. Menghitung nilai *Information Gain (IG)* pada setiap fitur dengan persamaan berikut :

$$IG\ (S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|s_i|}{|s|} x\ Entropy\ (S_i) \quad (4)$$

Dimana *S* adalah kumpulan dataset seluruhnya, *A* fitur pada Subset, *N* banyaknya partisi yang terdapat pada fitur A, | S_i | banyaknya subset dengan fitur A di partisi ke-I dan | *S* | adalah banyaknya kasus dalam dataset.

2.3.2. Gain Ratio

Metode seleksi fitur *Gain Ratio* adalah hasil pengembangan dari metode seleksi fitur *Information Gain*. Untuk menghitung nilai *Gain Ratio* dibutuhkan perhitungan *Information Gain* terlebih dahulu menggunakan persamaan (4) karena *Gain Ratio* adalah peningkatan perhitungan dari *Information Gain* [17]. Berikut ini adalah langkah perhitungannya:

a. Menghitung *Split Information* untuk setiap fitur dengan persamaan:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|}\ x\ log_2\left(\frac{D_j}{D}\right) \quad (5)$$

Dimana *D* adalah kumpulan dataset seluruhnya, A fitur pada subset, *v* banyaknya partisi pada fitur A, | D_j | banyaknya subset dengan fitur A di partisi ke-j dan | *D* | banyaknya kasus yang terdapat pada dataset

b. Menghitung *Gain Ratio* pada masing-masing fitur dengan persamaan:

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{SplitInfo\ (A)} \quad (6)$$

2.3.3. Gini Index

Gini index adalah probabilitas dua data yang dipilih secara acak yang memiliki kelas yang berbeda. *Gini index* digunakan untuk menghasilkan pohon klasifikasi dalam *Decision Tree*. Misalkan S adalah 1 kumpulan data bilangan s. Data ini memiliki jumlah kelas m yang berbeda (Ci, i = 1, ..., m). Berdasarkan kelasnya S dapat dibagi menjadi sebuah bilangan dari m himpunan bagian (Si, i = 1, ..., m), misal Si adalah sekumpulan data yang tergabung dalam kelas Ci dan si adalah jumlah data dari Si, maka *Gini Index* dapat dirumuskan sebagai berikut [10]:

$$Gini\ Index\ (S) = 1 - \sum_{i=1}^{m} \left(\frac{S_i}{S}\right)^2 \quad (7)$$

2.4. Algoritma K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) merupakan salah satu algoritma yang banyak digunakan untuk mengklasifikasikan data. Metode ini mengklasifikasikan data berdasarkan perhitungan jarak terdekat dari data uji ke data latih dan untuk menentukan kelas data baru menggunakan sistem *vote majority* dari jumlah K tetangga terdekat [3]. Tahapan klasifikasi data menggunakan algoritma K-NN dijelaskan dalam langkah-langkah sebagai berikut [12]:

a. Preprocessing dataset.
b. Membagi dataset menjadi dua bagian menjadi data latih dan data uji. Pada penelitian ini ditentukan data latih sebesar 70% dan data uji sebesar 30%.
c. Penentuan nilai K tetangga terdekat
d. Menghitung jarak antara data uji dan data latih menggunakan model jarak Euclidean dengan persamaan berikut:

$$D\_euclidean\ (x,y) = \|x - y\|\_2 = \sqrt{\sum_{j=1}^{N} |x - y|^2} \quad (8)$$

e. Mengurutkan data berdasarkan jarak terdekat sebanyak K
f. Penentuan kelas data uji berdasarkan kelas data pelatihan terdekat

Dalam proses klasifikasi data menggunakan algoritma K-NN pada penelitian ini ditentukan nilai K=1 sampai K=15. Proses klasifikasi dilakukan sebanyak 15 kali dengan nilai K yang berbeda. Selanjutnya hasil akurasi klasifikasi data pada K-NN tanpa seleksi fitur, K-NN dengan seleksi fitur *Information Gain*, *Gain Ratio* dan *Gini Index* akan dicatat dalam tabel perbandingan hasil akurasi kemudian dilakukan tahapan analisa dari hasil tersebut..

3. HASIL DAN PEMBAHASAN

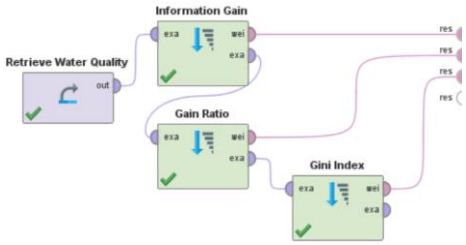
Hasil penelitian diperoleh dari pengujian dataset *Water Quality* menggunakan metode yang telah dibahas sebelumnya. Tahap awal pengujian dilakukan pada proses seleksi fitur untuk memilih fitur-fitur yang relevan. Setelah diperoleh hasil seleksi fitur selanjutnya dilakukan pengujian kinerja algoritma K-NN dalam proses klasifikasi data menggunakan seleksi fitur tersebut. Untuk perhitungan bobot fitur penulis menggunakan aplikasi *Rapid Miner Studio* sedangkan untuk menguji kinerja algoritma K-NN menggunakan bahasa pemrograman *Python* yang dijalankan melalui *Google Colaboratory*.

Hasil penelitian secara rinci dibahas pada bagian di bawah ini, baik dari hasil seleksi fitur hingga hasil akurasi pada proses klasifikasi data.

3.1. Hasil Seleksi Fitur

Tahapan pengujian diawali dengan mengolah dataset dengan mengganti data karakter menjadi numerik agar dapat dilakukan proses perhitungan bobot dan jarak. Selanjutnya dilakukan proses seleksi fitur dengan menghitung nilai bobot pada masing-

masing fitur. Pada pengujian ini dapat diketahui fitur-fitur yang memiliki pengaruh terbesar dan terendah pada dataset dengan menghitung bobot fiturnya. Dalam proses perhitungan bobot fitur untuk *Information Gain*, *Gain Ratio* dan *Gini Index* pada aplikasi *Rapid Miner Studio* menggunakan model berikut :



Gambar 2. Model Perbandingan Bobot Fitur

Hasil perhitungan bobot dari model di atas selanjutnya dicatat dalam tabel perbandingan hasil bobot dan untuk menentukan proporsi fitur berdasarkan nilai bobotnya dilakukan perhitungan menggunakan persamaan (2) seperti contoh berikut :

Total bobot fitur *Information Gain* = 3,891
Bobot fitur aluminium =1, sehingga diperoleh
$\text{Proporsi}_{(\text{aluminium})} = \frac{1}{3{,}891}\ x\ 100\% = 26\%.$
Total bobot fitur *Gain Ratio* = 3,116
Bobot fitur aluminium =0,56, sehingga diperoleh
$\text{Proporsi}_{(\text{aluminium})} = \frac{0{,}56}{3{,}116}\ x\ 100\% = 18\%.$
Total bobot fitur *Gini Index* = 3,383
Bobot fitur aluminium =1, sehingga diperoleh
$\text{Proporsi}_{(\text{aluminium})} = \frac{1}{3{,}383}\ x\ 100\% = 30\%.$

Proses perhitungan proporsi dilakukan pada semua fitur untuk setiap metode seleksi fitur sehingga diperoleh hasil secara rinci seperti yang terlihat pada tabel di bawah ini :

Tabel 2. Perbandingan hasil perhitungan bobot dan proporsi fitur pada dataset *Water Quality*

No	Fitur	Information Gain		Gain Ratio		Gini Index	
		Bobot	Proporsi	Bobot	Proporsi	Bobot	Proporsi
1	aluminium	1,000	26%	0,560	18%	1,000	30%
2	ammonia	0,007	0%	0,065	2%	0,006	0%
3	arsenic	0,395	10%	0,174	6%	0,305	9%
4	barium	0,084	2%	0,088	3%	0,069	2%
5	cadmium	0,776	20%	0,368	12%	0,682	20%
6	chloramine	0,490	13%	0,222	7%	0,390	12%
7	chromium	0,395	10%	0,176	6%	0,328	10%
8	copper	0,047	1%	0,065	2%	0,036	1%
9	flouride	0,000	0%	0,021	1%	0,000	0%
10	bacteria	0,004	0%	0,000	0%	0,003	0%
11	viruses	0,072	2%	0,030	1%	0,059	2%
12	lead	0,010	0%	0,019	1%	0,007	0%
13	nitrates	0,038	1%	1,000	32%	0,032	1%
14	nitrites	0,103	3%	0,072	2%	0,075	2%
15	mercury	0,008	0%	0,001	0%	0,007	0%
16	perchlorate	0,250	6%	0,109	4%	0,205	6%
17	radium	0,050	1%	0,072	2%	0,040	1%
18	selenium	0,006	0%	0,008	0%	0,005	0%
19	silver	0,102	3%	0,044	1%	0,087	3%
20	uranium	0,054	1%	0,022	1%	0,046	1%
	Total	3,891	100%	3,116	100%	3,383	100%

Nilai bobot dan proporsi fitur selanjutnya diurutkan dari nilai tertinggi sampai nilai terendah. Pemilihan fitur diambil dari proporsi fitur yang paling tinggi hingga pada fitur dengan jumlah total proporsi mencapai maksimal 90% seperti yang telah ditentukan. Untuk fitur lainnya yang tidak terpilih akan dibuang dari dataset.

Dataset *Water Quality* yang memiliki 20 fitur, setelah dihitung nilai bobot dan proporsinya menggunakan ketiga metode seleksi fitur diperoleh hasil yang berbeda. Namun pada hasil fitur yang dipilih terdapat kesamaan jika dilihat dari persentase proporsinya. Tabel di bawah ini merupakan hasil fitur yang terpilih menggunakan metode seleksi fitur *Information Gain*, *Gain Ratio* dan *Gini Index*.

Tabel 3. Hasil fitur yang terpilih pada dataset *Water Quality*

No	Metode seleksi fitur					
	Information Gain		*Gain Ratio*		*Gini Index*	
	Fitur	Proporsi	Fitur	Proporsi	Fitur	Proporsi
1	aluminium	26%	nitrates	32%	aluminium	30%
2	cadmium	20%	aluminium	18%	cadmium	20%
3	chloramine	13%	cadmium	12%	chloramine	12%
4	chromium	10%	chloramine	7%	chromium	10%
5	arsenic	10%	chromium	6%	arsenic	9%
6	perchlorate	6%	arsenic	6%	perchlorate	6%
7	nitrites	3%	perchlorate	4%	silver	3%
8	silver	3%	barium	3%		
9			nitrites	2%		
	Total	90%		89%		89%

3.2. Hasil Klasifikasi Data

Pengujian proses klasifikasi data dilakukan sebanyak empat kali pengujian pada dataset dengan nilai K yang ditentukan dari K=1 hingga K=15. Pengujian menggunakan dataset *Water Quality* yang dibagi secara langsung dengan ketentuan 70% data dijadikan sebagai data latih dan 30% data dijadikan sebagai data uji. Tahap awal yang dilakukan adalah menguji proses klasifikasi data dengan algoritma K-NN tanpa seleksi fitur pada dataset. Hasil yang diperoleh akan dicatat pada tabel sebagai hasil awal yang akan dibandingkan pada pengujian selanjutnya yaitu proses klasifikasi data menggunakan algoritma K-NN dengan seleksi fitur *Information Gain*, *Gain Ratio* dan *Gini Index.*

Klasifikasi data pada algoritma K-NN dengan seleksi fitur disesuaikan dengan fitur yang dipilih berdasarkan tabel 3 di atas yaitu *Information Gain* (IG) menggunakan 8 fitur yang terpilih, *Gain Ratio* (GR) menggunakan 9 fitur yang terpilih dan *Gini Index* (GI) menggunakan 7 fitur yang terpilih. Masing-masing hasil klasifikasi data dengan metode seleksi fitur tersebut juga dicatat hasilnya pada tabel. Dengan demikian dapat dilihat hasil perbandingan akurasi yang diperoleh dari klasifikasi dataset *Water Quality*.

Pengujian ini bertujuan untuk mengetahui metode seleksi fitur yang paling optimal dalam meningkatkan kinerja algoritma K-NN. Peningkatan kinerja dapat diketahui dari nilai akurasi klasifikasi data yang semakin meningkat. Nilai akurasi hasil

pengujian algoritma dapat diperoleh dari perhitungan menggunakan persamaan (1) seperti contoh berikut:
Jumlah prediksi benar = 2000
Jumlah seluruh data uji = 2400, sehingga diperoleh

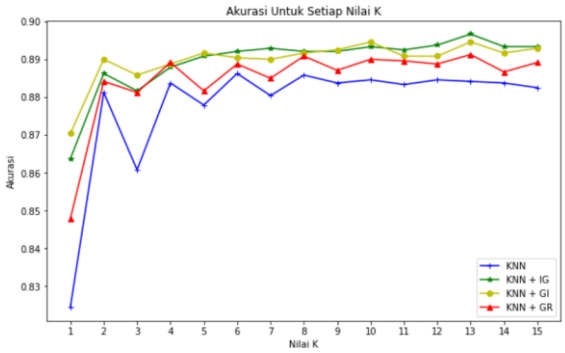
$$Akurasi = \frac{2000}{2400}\ x\ 100\% = 83{,}88\%$$

Tabel berikut ini menunjukkan perbandingan hasil akurasi pada algoritma K-NN dengan seleksi fitur yang berbeda pada setiap nikai K.

Tabel 4. Perbandingan hasil akurasi penggunaan seleksi fitur pada setiap nilai K

Nilai K	KNN	KNN + IG		KNN + GR		KNN + GI	
	Akurasi	Akurasi	Peningkatan	Akurasi	Peningkatan	Akurasi	Peningkatan
1	82,45%	86,37%	3,92%	84,79%	2,33%	87,04%	4,59%
2	88,12%	88,62%	0,50%	88,41%	0,29%	89,00%	0,88%
3	86,08%	88,16%	2,08%	88,12%	2,04%	88,58%	2,50%
4	88,37%	88,79%	0,42%	88,91%	0,54%	88,87%	0,50%
5	87,79%	89,08%	1,29%	88,16%	0,37%	89,16%	1,38%
6	**88,62%**	89,20%	0,58%	88,87%	0,25%	89,04%	0,42%
7	88,04%	89,29%	1,25%	88,50%	0,46%	89,00%	0,96%
8	88,58%	89,20%	0,63%	89,08%	0,50%	89,16%	0,58%
9	88,37%	89,20%	0,83%	88,70%	0,33%	89,25%	0,88%
10	88,45%	89,33%	0,88%	89,00%	0,54%	**89,45%**	1,00%
11	88,33%	89,25%	0,92%	88,95%	0,62%	89,08%	0,75%
12	88,45%	89,37%	0,92%	88,87%	0,42%	89,08%	0,63%
13	88,41%	**89,66%**	1,25%	**89,12%**	0,71%	**89,45%**	1,04%
14	88,37%	89,33%	0,96%	88,66%	0,29%	89,16%	0,79%
15	88,25%	89,33%	1,08%	88,91%	0,67%	89,29%	1,04%
Rata-Rata	87,78%	88,95%	1,17%	88,47%	0,69%	88,97%	1,19%

Selain dapat dilihat dari tabel di atas, hasil perbandingan nilai akurasi juga dapat dilihat pada grafik di bawah ini. Melalui grafik dapat dilihat dengan jelas bahwa semua metode seleksi fitur yang digunakan mampu meningkatkan nilai akurasi algoritma K-NN. Grafik untuk algoritma K-NN tanpa seleksi fitur berada di paling bawah dengan nilai akurasi terendah untuk semua nilai K yang ditentukan.



Gambar 3. Grafik perbandingan nilai akurasi pada setiap nilai K.

4. DISKUSI

Hasil perhitungan bobot dan proporsi fitur pada dataset *Water* Quality menggunakan metode seleksi fitur *Information Gain*, *Gain Ratio* dan *Gini Index* seperti yang ditunjukkan pada tabel 2 ternyata memiliki nilai bobot dan proporsi fitur yang berbeda-beda. Meskipun memiliki nilai yang berbeda namun memiliki persamaan pada hasil fitur yang dipilih. Pada penggunaan seleksi fitur *Information Gain* terpilih 8 fitur dengan total proporsi bobot 90% yaitu *aluminium*, *cadmium*, *chloramine*, *chromium*, *arsenic*, *perchlorate*, *nitrites* dan *silver*. Pada seleksi fitur *Gain Ratio* terpilih 9 fitur dengan total proporsi bobot 89% yaitu *nitrates*, *aluminium*, *cadmium*, *chloramine*, *chromium*, *arsenic*, *perchlorate*, *barium* dan *nitrites*. Sedangkan pada seleksi fitur *Gini Index* terpilih 7 fitur dengan total proporsi bobot 89% yaitu *aluminium*, *cadmium*, *chloramine*, *chromium*, *arsenic*, *perchlorate* dan *silver*.

Dari fitur-fitur yang terpilih dapat dikatahui fitur yang memberikan pengaruh sangat besar pada dataset. Hal ini terlihat dari fitur yang terpilih menggunakan ketiga metode seleksi fitur yang digunakan dalam dataset yaitu *aluminium*, *cadmium*, *chloramine*, *chromium*, *arsenic* dan *perchlorate*. Fitur tersebut terpilih disemua metode seleksi fitur yang digunakan dengan nilai bobot dan proporsi yang tidak jauh berbeda.

Selanjutnya pengujian metode klasifikasi data dengan menerapkan metode seleksi fitur pada dataset ternyata mampu meningkatkan kinerja algoritma K-NN. Hal ini terlihat dari nilai akurasi yang mengalami peningkatan di semua nilai K yang ditentukan yaitu dari K=1 hingga K=15. Klasifikasi data dengan algoritma K-NN tanpa menggunakan seleksi fitur memiliki rata-rata nilai akurasi sebesar 87,78%. Sedangkan proses klasifikasi data pada algoritma K-NN menggunakan seleksi fitur memiliki rata-rata nilai akurasi yaitu *Information Gain* sebesar 88,95%, *Gain Ratio* sebesar 88,47% dan Gini Index sebesar 88,97%. Dengan demikian rata-rata peningkatan nilai akurasi dari ketiga metode seleksi fitur yang digunakan yaitu *Information Gain* sebesar 1,17%, *Gain Ratio* sebesar 0,69% dan *Gini Index* sebesar 1,19%.

Sedangkan untuk nilai akurasi tertinggi yang diperoleh dari hasil pengujian metode seleksi fitur pada algoritma K-NN dapat dilihat pada tabel di bawah ini:

Tabel 5. Nilai akurasi tertinggi dari setiap seleksi fitur

Seleksi Fitur	Akurasi Tertinggi	Nilai K
Information Gain	89,66%	13
Gain Ratio	89,12%	13
Gini Index	89,45%	10 dan 13

Berdasarkan tabel di atas terlihat bahwa nilai akurasi tertinggi dari pengujian ketiga metode seleksi fitur yang digunakan pada algoritma K-NN yaitu pada seleksi fitur *Information Gain* sebesar 89,66% pada nilai K=13, diikuti oleh *Gini Index* sebesar 89,45% pada nilai K=10 dan K=13, serta yang terakhir *Gain Ratio* sebesar 89,12% pada nilai K=13.

Penerapan metode seleksi fitur *Information Gain*, *Gain Ratio* dan *Gini Index* pada algoritma K-NN dalam penelitian ini terbukti mampu meningkatkan kinerja algoritma K-NN. Metode seleksi fitur yang paling optimal untuk meningkatkan nilai akurasi algoritma K-NN dalam proses klasifikasi dataset *Water Quality* adalah seleksi fitur *Information Gain* dengan nilai K yang paling optimal adalah K=13.

5. KESIMPULAN

Berdasarkan hasil pengujian dan pembahasan dari metode seleksi fitur yang diujikan pada algoritma K-NN pada penelitian ini dapat disimpulkan bahwa penggunaan seleksi fitur *Information Gain, Gain Ratio* dan *Gini Index* pada algoritma K-NN terbukti mampu meningkatkan kinerja algoritma K-NN yang dapat dilihat dari peningkatan nilai akurasi di setiap nilai K yang ditentukan yaitu K=1 hingga K=15.

Rata-rata nilai akurasi yang diperoleh pada tahap pengujian algoritma K-NN tanpa seleksi fitur sebesar 87,78%. Sedangkan pada pengujian algoritma K-NN dengan seleksi fitur memiliki rata-rata nilai akurasi yaitu *Information Gain* sebesar 88,95%, *Gain Ratio* sebesar 88,47% dan *Gini Index* sebesar 88,97%. Terjadi peningkatan nilai akurasi yang diperoleh pada masing-masing metode seleksi fitur yaitu *Information Gain* sebesar 1,17%, *Gain Ratio* sebesar 0,69% dan *Gini Index* sebesar 1,19%.

Nilai akurasi tertinggi diperoleh pada seleksi fitur *Information Gain* dengan nilai akurasi sebesar 89,66% pada nilai K=13, diikuti oleh *Gini Index* sebesar 89,45% pada nilai K=10, K=13 dan yang terakhir *Gain Ratio* sebesar 89,25% pada K=13. Sehingga dapat diketahui metode seleksi fitur yang paling optimal untuk meningkatkan nilai akurasi algoritma K-NN dalam proses klasifikasi dataset *Water Quality* pada penelitian ini adalah metode seleksi fitur *Information Gain* dengan nilai K yang paling optimal adalah K=13.

Penelitian ini masih dapat dikembangkan kembali pada penelitian berikutnya dengan menerapkan metode seleksi fitur dalam proses klasifikasi data, baik menggunakan algoritma K-NN maupun algoritma klasifikasi lainnya menggunakan beberapa dataset yang memiliki karakteristik data yang berbeda. Dengan harapan dapat memberikan hasil penelitian yang lebih optimal.

DAFTAR PUSTAKA

[1] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status", *IEEE 6th International Conference on System Engineering and Technology (ICSET)*, pp. 137-141, October 3-4. 2016.

[2] I.A. Angreni1 , S.A. Adisasmita , M.I. Ramli dan S. Hamid, ”Pengaruh Nilai K Pada Metode K-Nearest Neighbor (K-NN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan”, Rekayasa Sipil, vol. 7, no. 2 , pp. 63-70, September. 2018.

[3] S.K. Lidya, O.S. Sitompul, and S. Effendi, “Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) Dan KNearest Neighbor (K-NN)”, Seminar Nasional Teknologi Informasi dan Komunikasi, pp. 1-8, 2015.

[14] A.C. Khotimah and E. Utami, “Comparison Naïve Bayes Classifier, K-Nearest Neighbor And Support Vector Machine In The Classification Of Individual On Twitter Account”, J. Tek. Inform. (JUTIF), vol. 3, no. 3, pp. 673-680, Jun. 2022.

[15] Z. Pan, Y. Wang, and W. Ku, “A New K-Harmonic Nearest Neighbor Classifer Based On The Multi-Local Means”, *Expert Systems With Applications*,67: 115-125, 2016.

[16] S. Samsani, “An RST based Efficient Preprocessing Technique for Handling Inconsistent Data”,*IEEE International Conference on Computational Intelligence and Computing Research*, 1-6, 2016.

[17] X. Zhang, Z. Shi, X. Liu, and X. Li, “A Hybrid Feature Selection Algorithm For Classification Unbalanced Data Preprocessing”, *International Conference on Smart Internet of Things*, 1-6, 2018.

[18] A.A. Nababan, O.S. Sitompul, and O. S, and Tulus, “Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio”, MECnIT, 1-6, 2018.

[19] I.T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, “Comparison Of Classification Algorithm And Feature Selection In Bitcoin Sentiment Analysis”, J. Tek. Inform. (JUTIF), vol. 3, no. 3, pp. 739-744, Jun. 2022.

[20] T. Setiyorini and R. Asmono, “Implementation Of K-Nearest Neighbor And Gini Index Method In Classification Of Student Performance”, *techno*, vol. 16, no. 2, pp. 121-126, Sep. 2019.

[21] A.K. Ginting, M.S. Lydia, E.M. Zamzami, “Peningkatan Akurasi Metode K-Nearest Neighbor dengan Seleksi Fitur Symmetrical Uncertainty”, Jurnal Media Informatika Budidarma, vol. 5, no. 4, pp. 1714-1719, Okt. 2021.

[22] Y.A. Pratama, Tulus, and S. Effendi, “Selection Features To Improve The Accuracy Of K-Nearest Neighbor”, *EPRA International Journal of Research and Development (IJRD*), Vol. 4, pp. 115-119, 2019.

[23] J. Han, J. Pei, and M. Kamber, "Data Mining Concept and Techniques, 3rd edition," Morgan Kaufmann-Elsevier. vol. 2, no. 1, pp. 88-97, 2012.

[24] M. Hafidzullah, Sutrisno and Marji, “Seleksi Fitur dengan Information Gain pada Identifikasi Jenis Attention Deficit Hyperactivity Disorder Menggunakan Metode Modified K-Nearest Neighbor”, Jurnal Pengembangan Teknologi Informasi

dan Ilmu Komputer, vol. 3, no. 11, p. 10444-10452, Jan 2020.

[25] S. Chormunge, and S. Jena, "Efficient Feature Subset Selection Algorithm for High Dimensional Data", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 4, pp. 1880-1888, August 2016.

[26] A. Nermeen, Shaltout, M. El-Hefnawi, A. Rafea, and A. Moustafa, "Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts", *Proceedings of the World Congress on Engineering*, vol. 1, p.625-631, July. 2-4, 2014.

[27] G.F. Grandis, Y. Arumsari, dan Indriati, "Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 5, no. 8, pp. 3507-3514, Agustus 2021.