EMITOR: Jurnal Teknik Elektro

Klasifikasi Kualitas Air Sungai Berbasis Teknik Data Mining Dengan Metode K-Nearest Neighbor (K-NN)

Bayu Prihambodo*, Adam Wildan F Y, Eko Prayoga, Ahmad Jaffar
Jurusan Teknik Elektro/Fakultas Teknik – Universitas Negeri Malang
Malang, Indonesia
*bayu.prihambodo.1905366@students.um.ac.id

Abstract– *The river is a source of water that plays an essential role in life. River water quality should be maintained to avoid water pollution by industrial waste. In addition, the impact of polluted air on humans is spreading disease germs. Potagen content in contaminated water that is dangerous includes cholera, giardia, and typhoid bacteria, as well as other harmful viruses that can cause hepatitis and polio. Therefore it is crucial to maintain the quality of river water. Air quality maintenance is essential permanently to preserve air quality, considering its importance for the life of living things. So we need a way to classify the air based on its quality. One way to speed up the classification of IKA based on the data and parameters that are already available is the data mining process. Classification techniques help predict the target class at each data point. In this study, the data mining method is the K-Nearest Neighbors method because it can be applied to classifying river water quality with a total sample of 216 samples with seven parameters of the Water Quality Index (IKA) and four classes. Then the best results were obtained in the classification process using the KNN method, namely getting an accuracy of* 78.46% *with training data values and testing data, respectively* 70% *and* 30%*, and a value of* $k = 15$.

Abstrak– Sungai merupakan sumber air yang sangat berperan penting dalam kehidupan. Air sungai seharusnya dijaga kualitasnya sehingga terhindar dari pencemaran air oleh limbah industri. Selain itu, dampak dari air yang tercemar pada manusia, yaitu menyebarkan bibit penyakit. Kandungan patagen pada air berpolutan yang berbahaya termasuk bakteri kolera, giardia, dan tifus serta virus berbahaya lain yang dapat menyebabkan hepatitis dan polio. Sehingga menjaga kualitas air sungai merupakan hal sangat penting. Penentuan kualitas air sangat diperlukan untuk selalu menjaga kualitas air mengingat pentingnya manfaatnya untuk kehidupan makhluk hidup. Maka diperlukan suatu cara untuk mengklasifikasikan air tersebut berdasarkan kualitas. Salah satu cara yang dilakukan untuk mempercepat klasifikasi IKA berdasarkan data dan parameter yang telah tersedia yaitu dengan proses data mining. Teknik klasifikasi dapat membantu dalam memprediksi kelas target pada setiap titik data. Pada penelitian ini metode data mining yang digunakan algoritme *K-Nearst Neighbor* dapat digunakan sebagai alternatif pengelompokan kualitas air sungai menjadi 4 kelas dengan sampel sebanyak 216 sampel 7 parameter Indeks Kualitas Air (IKA) dan 4 kelas. Maka diperoleh hasil terbaik pada proses klasifikasi dengan metode KNN yaitu memperoleh akurasi sebesar 78,46% dengan nilai data training dan data testing masing-masing 70% dan 30%, serta nilai $k = 15$.

Kata Kunci– *polutan-river; water quality; metode K-Nearest Neighbor; klasifikasi IKA.*

I. PENDAHULUAN

SUNGAI merupakan sumber air yang memegang peranan penting dalam kehidupan. Sungai memegang peranan penting dalam proses pertanian dan perkebunan bahkan untuk sarana pembudidayaan ikan air tawar dan juga sarana rekreasi. Masyarakat Indonesia juga masih menangkap ikan untuk dikonsumsi. Namun tidak sedikit yang mencemari. Kualitas air sungai berubah akibat aktivitas manusia di lingkungan sungai. Beberapa pencemaran sungai disebabkan baik oleh organisme yang ada di sekitar sungai maupun oleh ulah manusia sebagai pengguna sungai.

Naskah diterima 13 Desember 2022, revisi 9 Maret 2023, terbit online 23 Maret 2023. `Emitor` merupakan Jurnal Teknik Elektro – Universitas Muhammadiyah Surakarta yang terakreditasi dengan Sinta 3 beralamat di `https://journals2.ums.ac.id/index.php/emitor/index`.

Efek dominan yang sangat terlihat dari tingkat pencemaran air adalah kerusakan yang diakibatkan oleh cara hidup masyarakat yang memanfaatkan alam. Limbah industri tentunya dibuang ke sungai pada beberapa bantaran sungai yang berdekatan dengan kawasan industri. Hal ini tentunya dapat mengakibatkan penurunan kualitas air sungai.

Limbah industri atau limbah lain yang dibuang ke perairan sungai tentunya akan berdampak pada air berpolutan mulai dari lingkungan yaitu dapat menyebabkan kematian pada hewan. Merebaknya alga yang ada di danau dan lingkungan laut karena mendapat nutrisi sehingga menghabiskan kadar oksigen pada air. Kesuburan tanah yang rendah disebabkan oleh air yang mengandung logam berat, asam sulfat, merkuri dan bahan kimia lainnya mengubah kandungan unsur hara tanah sehingga menjadi sulit bagi tanaman untuk hidup.

Dampak selanjutnya adalah pada manusia, yaitu menyebarkan bibit penyakit. Air yang tercemar mengandung beberapa polutan seperti kolera yang dapat menyebabkan hepatitis dan polio [1]. Selain itu pencemaran air bisa terlarut zat radioaktif yang dapat mengganggu kesehatan. Unsur logam, pestisida, dan merkuri dalam air yang terkontaminasi masuk ke dalam tubuh hewan, dicerna oleh tumbuhan, dan menjadi sumber makanan bagi manusia yang tidak layak untuk dikonsumsi.

Direktorat Jenderal Pengendalian Pencemaran dan Kerusakan Lingkungan Hidup dan Kehutanan (KLHK) mengatakan bahwa kondisi sebagian besar sungai di Indonesia tercemar berat [2]. Dengan demikian dalam mengendalikan pencemaran air pemerintah mengeluarkan Peraturan Pemerintah Nomor 82 Tahun 2001 sebagai bentuk upaya pengawasan dan pengendalian terhadap air pencemaran air sungai secara berkala demi menjaga kualitas dan kuantitas air sungai.

Berdasarkan PP No. 22 Tahun 2021 tentang Klasifikasi Kelas Air dibagi menjadi 4 kelas Kelas 1 adalah air yang dapat digunakan sebagai air baku dan digolongkan sebagai air minum. Kelas 2 adalah perairan untuk prasarana/sarana rekreasi, pembudidayaan ikan air tawar, peternakan, dan pengairan tanaman. Kelas 3 dapat digunakan untuk budidaya ikan air tawar, pembibitan dan pengairan tanaman. Kelas terakhir adalah kelas 4, yang digunakan untuk sistem irigasi dalam budidaya tanaman. Masing-masing dari kelas ini tentunya mempunyai nilai standar untuk mengklasifikasikan berada dimana kualitas air sungai tersebut. IKA (*Indeks Kualitas Air*) dapat ditentukan oleh beberapa parameter yaitu derajat keasaman (pH), konsentrasi TSS (*total suspended solid*), konsentrasi DO, konsentrasi BOD, konsentrasi COD, konsentrasi total Phospat, konsentrasi *Fecal Coliform*, konsentrasi Nitrat ($NO_3 - N$) [3].

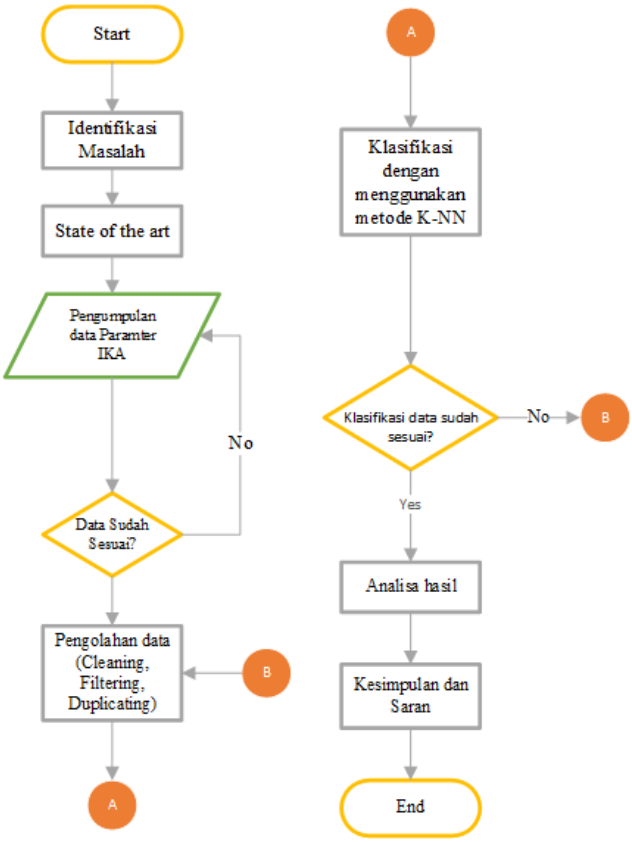
Penentuan kualitas air sangat diperlukan untuk selalu menjaga kualitas air mengingat pentingnya manfaatnya untuk kehidupan makhluk hidup. Maka diperlukan suatu cara untuk mengklasifikasikan air tersebut berdasarkan kualitasnya. Salah satu cara yang dilakukan untuk mempercepat klasifikasi IKA berdasarkan data dan parameter yang telah tersedia yaitu dengan proses data mining. Untuk memprediksi setiap titik data pada kelompok target, dapat menggunakan teknik klasifikasi. Terdapat beberapa metode dalam klasifikasi data mining, yaitu seperti Naive Bayes, K-Nearest Neighbor (K-NN), dan *Decision Trees*. Berbagai metode klasifikasi tersebut dapat menghasilkan nilai akurasi yang beragam. Pada penelitian ini sistem klasifikasi data menggunakan algoritme K-NN [4–9]. Klasifikasi bertujuan untuk menentukan kelas dari kualitas air sungai. Data penelitian didapatkan dari data pengecekan air dinas lingkungan hidup. Dimana proses klasifikasi akan menggunakan parameter-parameter indeks kualitas air (IKA) untuk menentukan kelas air sungai.

II. Metode Penelitian

Penelitian dilakukan di PDAM Tirta Kota Malang. Data diambil dari tahun 2019 sampai 2021 di 3 titik bagian sungai yang berada di Kota Malang. Pengambilan data pada 3 titik sungai tersebut meliputi Waduk Sutami Hilir 0,3 m, Cangkir Tambangan, dan Muara Kali Tengah. Pada penelitian ini data yang digunakan meliputi data kuantitatif dan data deskriptif. Data kuantitatif diperoleh dari perhitungan dan hasil simulasi, sedangkan data deskriptif diperoleh langsung dari pihak PDAM Kota Malang dengan melakukan pengamatan menggunakan alat pada tahun 2021. Penelitian ini akan mengolah data yang sudah ada dan tidak mencari kekurangan data. Berdasarkan parameter indeks kualitas air menurut PP No 22 Tahun 2021 peneliti mengurangi parameter konsentrasi total phospat, konsentrasi *fecal coliform*, dan konsentrasi nitrat($NO_3 - N$). Data yang diperoleh selanjutnya diolah menggunakan algoritma K-NN (K-Nearst Neighbor). Selanjutnya dilakukan uji validasi keakuratan metode yang digunakan dengan menambahkan *tools* RapidMiner Studio yaitu performa di akhir setelah K-NN. Sebeum dilakukan klasifikasi dilakukan pembagian data dengan *tools split* data untuk membagi data menjadi data testing dan data *training*, dimana data testing akan diolah menggunakan algoritma K-NN (K-Nearst Neighbor) agar dapat diperoleh hasil klasifikasi. Untuk alur proses penelitian dijelaskan pada Gambar 1 berikut :

Pada Gambar 1 sebelum memasukkan data untuk proses klasifikasi dilakukan beberapa pemrosesan data. Tujuan dari filtering data ini untuk memperoleh data yang tidak tumpang tindih atau data yang kurang. Pemrosesan data ini meliputi:

1. *Data Cleaning* adalah proses pembersihan data dengan cara menghilangkan noise, menghilangkan duplikasi data, memeriksa data yang tidak stabil, menghilangkan duplikasi data, menutup data yang kosong.
2. Data integrasi diperlukan pada pre-processing untuk menggabungkan beberapa data yang dibutuhkan pada data mining, mengurangi, menghilangkan pengulangan data, serta meniadakan hasil data set yang yang tidak sesuai. Sehingga dapat menambah akurasi dan kecepatan proses data mining.
3. Transformasi Data merupakan proses pengubahan data dan mengkonsolidasikan data, menjadi bentuk yang sesuai sehingga proses data mining lebih efisien dan pola yang ditemukan



Gambar 1: Flowchart alur penelitian

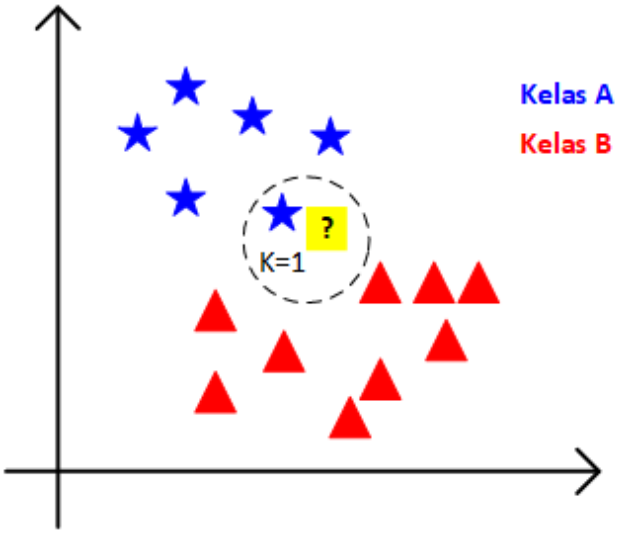
dapat lebih mudah dimengerti.

i. Metode K-Nearest Neighbors (K-NN)

KNN merupakan algoritme pembelajaran mesin terawasi yang digunakan untuk proses klasifikasi dan regresi [10]. Algoritme ini memanipulasi data *training* dan mengklasifikasikan data testing baru berdasarkan metrik jarak. Metode ini kemudian menemukan K-tetangga terdekat ke data uji, dan kemudian klasifikasi dilakukan oleh sebagian besar label kelas [11]. Algoritme ini masuk dalam kelas *instance based learning* dan juga *lazy learning* [12].

K-NN dibangun dengan mencari kelompok K objek pada data pelatihan yang paling dekat dengan objek dengan data yang dicari dan data testing. Algoritme K-NN bekerja dari metrik jarak terpendek dari *instance query* ke data testing untuk menentukan K-NN [13]. Untuk menentukan jara tetangga terdekat dan tetangga jauh dari suatu objek yang belum diketahui dihitung dengan menggunakan metode Euclidean distance. Euclidian distance yang merupakan ukuran jarak yang udah banyak digunakan untuk menghitung jarak data uji terlatih [14]. Perhitungan nilai jarak euclidian distance ada pada Persamaan (1).

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (1)$$



Gambar 2: Ilustrasi pengaruh nilai k pada metode K-NN

dengan $d(x,y)$ = Euclidean Distance, x_k = data *training* yang diperlukan, y_k = data tang akan dicari tetangga terdekatnya, k = *record* baris ke$-k$ dari tabel, dan n = total data *training*.

Penentuan nilai k yang optimal untuk metode *KNN* ini berhubungan dengan nilai dan jumlah data yang akan diklasifikasikan. Semakin tinggi nilai k yang digunakan maka akan lebih tinggi mengurangi dampak noise pada klasifikasi, tetapi mengaburkan batas antar klasifikasi. Dam proses penentuan nilai k dapat dilakukan mengoptimalan parameter seperti validasi silang.

ii. Pengujian Confusion Matrix

Setelah didapatkan output dari RapidMiner studio selanjutnya dibutuhkan pula pengukuran untuk menilai seberapa akurat klasifikasi yang dipilih dalam estimasi hasil dari klasifikasi tersebut menggunakan *confusion matrix*. Confusion matrix dibuat untuk mentabulasi kinerja setiap pengklasifikasi. Tabel confusion matrix akan memberikan informasi mengenai kelas yang ditentukan sebagaimana yang ada pada Tabel 1.

Tabel 1: Respons variabel confusion matriks

Class	Positive	Negative
Positive	True Positive (TP)	False Negatif (FN)
Negative	False Positive (FP)	True Negatif (TN)

Perhitungan prediksi akan ditampung pada confusion matrix. Berdasarkan nilai yang diperoleh dari confusion matrix, ditemukan nilai hasil “akurasi”, ”presisi”, ”recall”, dan ”F1-score” untuk mengevaluasi kinerja metode klasifikasi apapun [13]. Untuk menentukan performa dari klasifikasi dilakukan dengan

memasukkan parameter ke beberapa rumus. Untuk mencari akurasi menggunakan Persamaan (2), presisi menggunakan Persamaan (3), recall menggunakan Persamaan (4), dan F1-score menggunakan Persamaan (5).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1 - Score} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad (5)$$

iii. RapidMiner Studio

RapidMiner studio merupakan perangkat lunak berbasis *open-source* (bersifat terbuka) yang memberikan Solusi untuk menganalisis penambangan data, penambangan teks, dan analitis prediksi [15]. RapidMiner Studio menggunakan berbagai teknik deskriptif dan prediksi untuk memberikan informasi kepada pengguna sehingga mereka dapat membuat keputusan terbaik. RapidMiner studio mempunyai sistem input, output, visualisasi data, dan operator yang dapat dipakai secara langsung dan open source [16].

III. Hasil Penelitian dan Diskusi

Penelitian ini menggunakan 216 *record* data yang di peroleh dari di PDAM Kota Malang. Data yang sudah didapat kemudian dilakukan proses pemfilteran untuk memperoleh data yang akurat. Data kemudian diproses untuk diolah sesuai parameter-parameter yang terdapat pada PP Nomor 22 Tahun 2021 tentang penyelenggaraan Perlindungan dan Pengelolaan lingkungan Hidup. Analis didasarkan pada indeks kualitas air dimana klasifikasi dilakukan untuk menentukan kelas setiap sampel air. Variabel respons yang digunakan dalam penelitian ini dijelaskan pada Tabel 2 di bawah ini.

Data pada Tabel 2 di atas kemudian digunakan untuk menghitung jarak masing-masing sampel untuk menentukan jarak terdekat. Perhitungan jarak pada penelitian ini menggunakan metode Euclidean distance.

i. Pre-Prosessing Data

Data yang diperoleh pada penelitian ini perlu dilakukan proses pre-processing. Kegunaan dari pre-prosessing ini untuk memilah data. Atribut yang diberi label dalam penelitian ini adalah kelas air. Terdapat 4 kelas

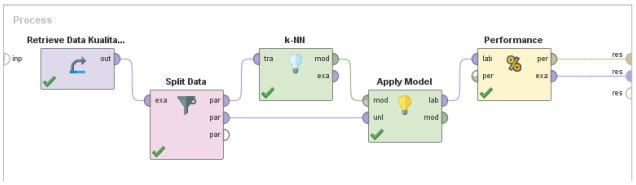
Tabel 2: Respons perilaku variabel terhadap metode jarak

Parameter	Variabel
Kelas kualitas sampel air (Kelas 1, Kelas 2, Kelas 3, Kelas 4)	Y
Temperatur	x_1
Padatan Teruspensi Total (TSS)	x_2
pH	x_3
Warna	x_4
BOD	x_5
COD	x_6
DO	x_7

yang akan diberikan label yaitu kelas 1, kelas 2, kelas 3, dan kelas 4. Proses pemberian label dapat dilakukan pengaturan warna pada label agar mempermudah proses penelitian. Metode yang digunakan untuk preprocessing pada penelitian ini yaitu validasi data untuk mendapatkan data yang baik dengan akurasi yang akurat, dilakukan peninjauan kembali kelengkapan data dan mengidentifikasi jenis data agar mendapatkan hasil yang akurat. Selanjutnya dilakukan validasi data yang non konsisten, dan data yang hilang, dimana dari kondisi data yang mentah menjadi data yang siap diolah dan dapat dianalisis karena proses *cleaning* data dan *filtering* data pada proses validasi data [17].

ii. Data Training dan Data Testing

Proses pembagian data langsung menggunakan operator yang ada di RapidMiner studio. Pada penelitian ini desain proses yang dilakukan pada software RapidMiner studio ditunjukkan pada Gambar 3.



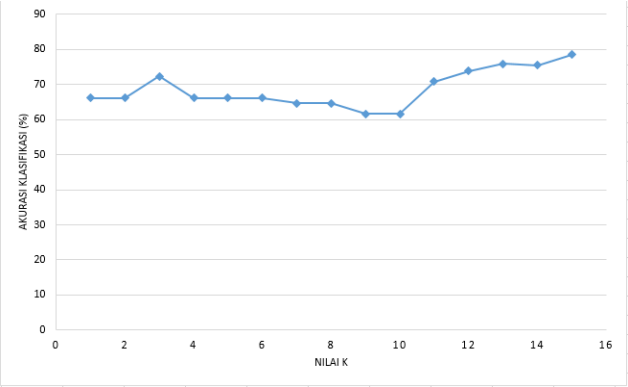
Gambar 3: Alur pemrosesan data menggunakan metode KNN pada software *RapidMiner*

Operator *split* data berguna untuk melakukan pembagian data yang telah di inputkan ke dalam RapidMiner. Data akan terbagi menjadi data training dan data testing, yang rasio dari data *training* sebesar 0.7 (70%) dan data testing 0.3(30%), sehingga untuk data *training* sebanyak 151 data dan data testing sebanyak 65 data.

iii. Pengujian Pengaruh Nilai k

Dengan memperhatikan nilai k yang digunakan pada pengujian kali ini sebanyak 12, dimulai dari 1, 2, 3,

4, 5, 6, 7, 8, 9, 10, 11, 12 dan 13 pada uji coba yang dilakukan menggunakan RapidMiner sebanyak 3 kali *running*. Pengujian ini dilakukan menggunakan parameter indeks kualitas air dengan rasio data *training* dan data *setting* masing-masing adalah 0.7:0.3 dengan secara acak pengambilan datanya, yaitu data *training* yang digunakan sebanyak 70% dari *datasheet* dan untuk data uji 30% dari data sampel kualitas air sungai yang dipilih secara acak. Kemudian setelah dilakukan pengujian data hasil uji nilai *k* terhadap akurasi klasifikasi air sungai didapatkan hasil seperti pada Gambar 4.



Gambar 4: Grafik hasil pengujian pengaruh nilai *k* pada klasifikasi air sungai

Grafik pada Gambar 4 dapat terlihat bahwa akurasi paling tinggi pada pengujian saat nilai $k = 15$ dengan memperoleh akurasi $78,46\%$. Untuk nilai dengan asil akurasi terendah pada pengujian ini yaitu saat nilai $k = 9$. Dari grafik di atas dapat disimpulkan bahwa nilai k sangat berpengaruh terhadap nilai output akurasi yang dihasilkan. Nilai akurasi ini dipengaruhi banyak faktor. Nilai validitas yang rendah serta perbandingan data yang semakin banyak menyebabkan error dalam proses klasifikasi. Selain itu label klasifikasi yang salah akan menurunkan tingkat akurasi pada saat klasifikasi. Hasil rata-rata pengujian nilai k menghasilkan akurasi sebesar $68,7\%$.

iv. Perhitungan Multiple Confusion Matrix

Confusion Matrix adalah metode klasifikasi berdasarkan hasil klasifikasi yang dilakukan, dan akurasi klasifikasi mempengaruhi kinerja klasifikasi [18]. Selain itu, Confusion matrix sebagai informasi komparatif tentang proses klasifikasi oleh sistem (model) dan klasifikasi hasil yang sebenarnya. Confusion matrix memberikan perhitungan nilai precision, recall, dan accuracy [19]. Multi confusion matrix digunakan pada penelitian ini dikarenakan *labeling* lebih dari 2 kondisi [20]. Klasifikasi data dengan algoritme K-Nearest neighbors menghasilkan Confusion Matrix KNN dalam Tabel 3 berikut.

Tabel 3: Hasil Confusion Matrix dengan metode $K - NN$ pada *T.K = true class* dan *P.K = prediction class*

Prediksi Kelas	*T.K* 2	*T.K* 3	*T.K* 1	*T.K* 4
P.K 1	0	0	0	0
P.K 2	26	5	2	0
P.K 3	3	22	0	5
P.K 4	0	1	0	1

Pada Tabel 3 dapat dilihat bahwa total nilai *True Positive* (TP) sebanyak 43, merupakan nila yang sesuai dengan klasifikasi air sesuai dengan kelasnya. Nilai *False Positif* total (FP) sebanyak 22. Untuk menghitung nilai accuracy total maka dilakukan perhitungan nilai F1-Score. F1-Score merupakan perbandingan antara *recall* dan *precision*. Tabel 4 berikut merupakan hasil dari masing-masing kelas.

Tabel 4: Hasil perhitungan $F1-$Score pada masing-masing kelas

Kelas	Precision	Recall	F1-Score
Kelas 2	0,7879	0,8966	0,83875
Kelas 3	0,7333	0,7857	0,7585
Kelas 1	0	0	0
Kelas 4	0,5	0,1667	0,25

Pada Tabel 4 nilai F1-Score digunakan untuk menentukan *accuracy* dari metode K-Nearest Neighbors dalam melakukan proses klasifikasi kualitas air dan didapatkan nilai *accuracy* sebesar 66,15%. Proses klasifikasi menggunakan metode K-Nearst Neighbors dapat bekerja dengan baik. Penggunaan RapidMiner studio Studio membuat proses semakin berjalan dengan cepat. Beberapa operator yang digunakan pada RapidMiner studio Studio diantaranya terdapat *Retrieve*, *Split Data*, Metode K-NN, *Apply Model*, dan *Performance*. Operator tersebut memiliki fungsi masing-masing untuk melakukan proses klasifikasi.

IV. KESIMPULAN

Berdasarkan hasil paparan di atas dan telah dilakukan analisis yang mendalam tentang penggunaan metode K-Nearst Neighbors untuk melakukan proses klasifikasi kualitas sungai, maka dapat diambil beberapa kesimpulan yaitu: Metode algoritme K-Nearst Neighbor dapat digunakan sebagai alternatif pengelompokan kualitas air sungai menjadi 4 kelas dengan sampel sebanyak 216 sampel 7 parameter Indeks Kualitas Air (IKA) dan 4 kelas. Hasil terbaik pada proses klasifikasi dengan metode

KNN yaitu memperoleh akurasi sebesar 78,46% dengan nilai data training dan data testing masing-masing 70% dan 30%, serta nilai $k = 15$. Nilai k sangat terpengaruh terhadap besarnya *confidence*, sehingga sangat mempengaruhi nilai klasifikasi *accuracy*.

DAFTAR PUSTAKA

[1] A. Tangkelayuk, "The klasifikasi kualitas air menggunakan metode knn, naïve bayes, dan decision tree," *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, vol. 9, no. 2, pp. 1109–1119, 2022.

[2] I. G. Vidiastanta, N. Hidayat, dan R. K. Dewi, "Komparasi metode k-nearest neighbors (k-nn) dengan support vector machine (svm) untuk klasifikasi status kualitas air," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, vol. 2548, p. 964X, 2020.

[3] R. A. Arnomo, "Implementasi algoritma k-nearest neigboruntukidentifikasi kualitas air (studi kasus: Pdam kota surakarta)," Ph.D. dissertation, STMIK Sinar Nusantara Surakarta, 2017.

[4] T. S. Wahyuni dan D. Kartikasari, "Analysis of well water quality based on physics, chemical, and microbiology parameters in iain tulungagung area," *Jurnal Akademika Kimia*, vol. 9, no. 4, pp. 245–250, 2020.

[5] A. D. Sutadian, N. Muttil, A. G. Yilmaz, dan B. Perera, "Using the analytic hierarchy process to identify parameter weights for developing a water quality index," *Ecological Indicators*, vol. 75, pp. 220–233, 2017.

[6] A. Sánchez, E. Cohim, dan R. Kalid, "A review on physicochemical and microbiological contamination of roof-harvested rainwater in urban areas," *Sustainability of Water Quality and Ecology*, vol. 6, pp. 119–137, 2015.

[7] J. C. Egbueri, "Water quality appraisal of selected farm provinces using integrated hydrogeochemical, multivariate statistical, and microbiological technique," *Modeling Earth Systems and Environment*, vol. 5, no. 3, pp. 997–1013, 2019.

[8] M. A. Nazir, A. Yasar, M. A. Bashir, S. H. Siyal, T. Najam, M. S. Javed, K. Ahmad, S. Hussain, S. Anjum, E. Hussain *et al.*, "Quality assessment of the noncarbonated-bottled drinking water: comparison of their treatment techniques," *International journal of environmental analytical chemistry*, vol. 102, no. 19, pp. 8195–8206, 2022.

[9] M. Lee, M. Kim, Y. Kim, dan M. Han, "Consideration of rainwater quality parameters for drinking purposes: A case study in rural vietnam," *Journal of environmental management*, vol. 200, pp. 400–406, 2017.

[10] A. Danades, D. Pratama, D. Anggraini, dan D. Anggriani, "Comparison of accuracy level k-nearest neighbor algorithm and support vector machine algorithm in classification water quality status," in *2016 6th International Conference on System Engineering and Technology (ICSET)*. IEEE, 2016, pp. 137–141.

[11] M. R. Nikoo, R. Kerachian, dan M. R. Alizadeh, "A fuzzy knn-based model for significant wave height prediction in large lakes," *Oceanologia*, vol. 60, no. 2, pp. 153–168, 2018.

[12] D. Cahyanti, A. Rahmayani, dan S. A. Husniar, "Analisis performa metode knn pada dataset pasien pengidap kanker payudara," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 39–43, 2020.

[13] T. Adithiyaa, D. Chandramohan, dan T. Sathish, "Optimal prediction of process parameters by gwo-knn in stirring-squeeze casting of aa2219 reinforced metal matrix composites," *Materials Today: Proceedings*, vol. 21, pp. 1000–1007, 2020.

[14] R. Huang, C. Cui, W. Sun, dan D. Towey, "Poster: Is euclidean distance the best distance measurement for adaptive random testing?" in *2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*. IEEE, 2020, pp. 406–409.

[15] A. N. Yuliarina dan H. Hendry, "Comparison of prediction analysis of gofood service users using the knn & naive bayes algorithm with rapidminer software," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 4, pp. 847–856, 2022.

[16] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, dan C. Lauw, "Penerapan aplikasi rapidminer untuk prediksi nilai tukar rupiah terhadap us dollar dengan metode regresi linier," *Jurnal Nasional Teknologi dan Sistem Informasi (TEKNOSI)*, vol. 7, no. 1, pp. 8–17, 2021.

[17] W. Xing dan Y. Bei, "Medical health big data classification based on knn classification algorithm," *IEEE Access*, vol. 8, pp. 28 808–28 819, 2019.

[18] S. Haghighi, M. Jasemi, S. Hessabi, dan A. Zolanvari, "Pycm: Multiclass confusion matrix library in python," *Journal of Open Source Software*, vol. 3, no. 25, p. 729, 2018.

[19] D. Krstinić, M. Braović, L. Šerić, dan D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Computer Science & Information Technology*, vol. 1, 2020.

[20] J. Görtler, F. Hohman, D. Moritz, K. Wongsuphasawat, D. Ren, R. Nair, M. Kirchner, dan K. Patel, "Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–13.