JUSTINDO
Jurnal Sistem dan Teknologi Informasi
http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO
E-ISSN: 25541-5735, P-ISSN: 2502-5724
Volume 7, No. 2, Agustus 2022

Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma *K-Nearest Neighbor* dan *Logistic Regression*

Baiq Nurul Azmi[1], Arief Hermawan[2], Donny Avianto[3]

[1,2]Magister Teknologi Informasi, Universitas Teknologi Yogyakarta
[3]Informatika, Universitas Teknologi Yogyakarta
Email: [1]6210211006.baiq@student.uty.ac.id, [2]ariefdb@uty.ac.id, [3]donny@uty.ac.id

(Naskah masuk: 10 Juli 2022, diterima untuk diterbitkan: 10 Agustus 2022, Terbit: 28 Agustus 2022)

ABSTRAK

Air bersih merupakan komponen penting untuk mendukung keberlangsungan hidup manusia. Perkembangan industri dan semakin bervariasinya aktivitas manusia berdampak pada penurunan kualitas air di area tersebut. Penurunan tingkat kualitas air dapat menyebabkan air menjadi tidak layak untuk dikonsumsi bahkan berbahaya untuk dikonsumsi. Kemampuan mengklasifikasi kualitas air secara akurat sangat diperlukan untuk menghindari penurunan tingkat kualitas air. Penelitian sebelumnya menunjukkan bahwa jumlah fitur yang digunakan untuk klasifikasi kualitas air sangat banyak. Jumlah fitur yang banyak ini memang dapat membantu metode pengklasifikasi untuk melihat domain permasalahan secara menyeluruh. Namun, belum ada penelitian yang meninjau secara detail apakah jumlah fitur yang banyak benar-benar diperlukan untuk mendapatkan hasil terbaik. Penelitian ini mengkaji penggunaan metode *principal component analysis (PCA)* untuk menemukan jumlah fitur yang paling optimal dalam konteks klasifikasi kualitas air. Penelitian ini menggunakan data kualitas air di lingkungan perkotaan yang diperoleh dari situs kaggle. Total data yang digunakan adalah 8000 baris data dengan 21 fitur untuk setiap baris data yang ada. Fitur hasil *principal component analysis* kemudian dijadikan input untuk dua metode klasfikasi yaitu *k-nearest neighbor (kNN)* dan *logistic regression*. Penggunaan dua metode klasifikasi yang berbeda ini bertujuan menemukan tingkat akurasi terbaik untuk data yang digunakan. Hasil eksperimen menunjukkan metode *k-nearest neighbor* mampu memberikan performa yang lebih baik dibandingkan *logistic regression* dengan pencapaian nilai akurasi 90.8%, presisi 90.0%, dan recall 91.0%. Hasil ini didapatkan dengan melibatkan seluruh fitur yang ada dan nilai k=9, sehingga dapat disimpulkan bahwa jumlah fitur yang banyak pada konteks klasifikasi kulitas air memang diperlukan untuk mendapatkan nilai akurasi yang tinggi.

Kata kunci: *kualitas, air, PCA, kNN, logistic, regression*

ABSTRACT

Clean water is an important part of human life. The water quality in some areas may be getting worse because industry is growing and people are doing more things. If the quality of the water goes down, it may become undrinkable or even hazardous to consume. It is very important to be able to accurately classify the water's quality to avoid the bad effects of using bad water. Previous study has demonstrated that a large number of features are required to accurately classify water quality. This large number of features can help the classifier approach understand the whole problem domain for classifying water quality. However, no research has ever examined whether a large number of

features are required for best performance. This study investigates the use of principal component analysis (PCA) to determine the optimal number of features for classifying water quality. This study uses data on water quality in urban environments obtained from the kaggle site. There are 8000 rows of data in total, with 21 features per row. The features generated by the principal component analysis are then fed into the k-nearest neighbor (kNN) and logistic regression. The purpose of comparing these two classification methods is to determine which one provides the highest level of accuracy. The experimental results demonstrate that the k-nearest neighbor technique outperforms logistic regression with an accuracy of 90.8%, precision of 90%, and recall of 91%. This best result is obtained by using all features and the value of k = 9, hence it can be inferred that a large number of features are required for accurate categorization of water quality.

Keywords: *water, quality, PCA, kNN, logistic, regression.*

1. PENDAHULUAN

Air adalah senyawa kimia yang sangat dibutuhkan bagi kelangsungan hidup makhluk hidup yang ada di bumi. Air bersih dan sanitasi dengan kondisi layak adalah kebutuhan dasar manusia. Salah satu poin dalam tujuan pembangunan berkelanjutan atau *Sustainable Development Goals* (SDGs) pada sektor lingkungan hidup adalah memastikan masyarakat mencapai akses universal air bersih dan sanitasi (SDGS, 2022).

Air rentan terkontaminasi oleh bakteri-bakteri dan zat mineral yang berbahaya bagi tubuh manusia. Hal tersebut dapat terjadi karena sumber air atau lingkungan sekitarnya telah tercemar. Air yang telah tercemar akan memberikan dampak pada akses air bersih atau air minum yang aman dikonsumsi oleh manusia. Beberapa faktor yang mempengaruhi akses air minum yang aman dikonsumsi antara lain adalah aktivitas manusia yang semakin tinggi dan proses industri yang semakin kompleks (Kustanto, 2020). Data yang dihimpun oleh Bank Dunia pada 2014, sekitar 780 juta orang tidak memiliki akses air bersih dan lebih dari 2 miliar penduduk bumi tidak memiliki akses terhadap sanitasi. Kekurangan akses terhadap air bersih mengakibatkan ribuan nyawa melayang tiap hari dan kerugian materi hingga 7 persen dari PDB dunia.

Penurunan kualitas air menyebabkan urgensi kebutuhan untuk mengawasi, menilai, dan mengklasifikasi kualitas air yang layak untuk konsumsi. Kualitas air yang sesuai dengan standar kesehatan dapat diketahui dari zat-zat atau mineral yang terkandung di dalamnya. Dataset tentang kualitas air pada situs *kaggle* memiliki banyak fitur terkait zat-zat kandungan air yang dijadikan parameter standar air bersih dan dapat dijadikan data untuk melakukan klasifikasi kualitas air.

Proses klasifikasi kualitas air dapat dilakukan dengan berbagai macam metode klasifikasi. Penelitian terkait klasifikasi kualitas air sebelumnya dilakukan oleh (Pritalia, 2022) yang bertujuan untuk mengklafisikasi air layak minum dengan dataset dari *water potability* dari kaggle yang memiliki 10 fitur yaitu *pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, dan Potablity*. Proses klasifikasi menggunakan algoritma *machine learning* yaitu *Decision Tree, Logistic Regression, Random Forest, SVM, Naïve Bayes, dan KNN*. Hasil penelitian menunjukkan dari enam algoritma yang digunakan, *Random Forest* memiliki performa akurasi yang paling optimal yaitu dengan akurasi latih sebesar 80% serta akurasi uji sebesar 85%.

Penelitian lain dilakukan oleh (Rahman, Hidayat and Supianto, 2018) untuk klasifikasi kualitas air bersih pada PDAM Tirta Kencana Kabupaten Jombang

p-ISSN : 2502-5724; e-ISSN : 2541-5735

dengan membandingkan metode *K-Nearest Neighbor* dan *Naïve Bayes* menghasilkan hasil akurasi yang cukup tinggi, yaitu untuk metode *K-Nearest Neighbor* memiliki akurasi 82,42% sedangkan metode *Naïve Bayes* memiliki akurasi 70,32%. metode *K-Nearest Neighbor* menghasilkan performa dan akurasi yang lebih baik untuk mengklasifikasi kualitas air bersih pada studi kasus tersebut, dibandingkan metode *Naïve Bayes.*

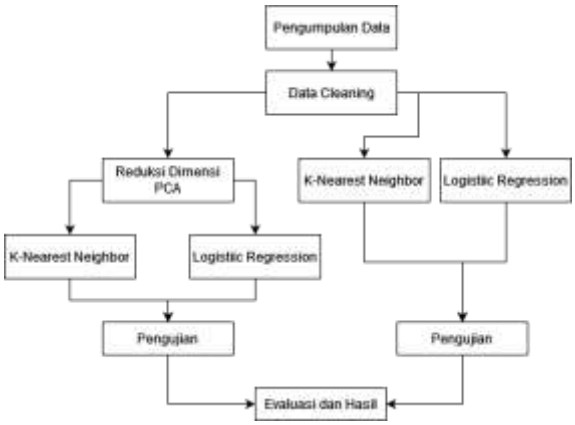
Faktor banyaknya fitur atau atribut untuk menentukan kualitas air dapat mempengaruhi kinerja model klasifikasi. Penggunaan seluruh fitur dalam proses klasifikasi dapat menurunkan performa klasifikasi dan memiliki waktu komputasi yang tinggi, sehingga pendekatan yang dapat dilakukan dengan melakukan reduksi dimensi.

Upaya untuk mengoptimalkan kinerja klasifikasi pada penelitian ini dengan melakukan beberapa pendekatan, yaitu penggunakan metode *Principal Component Analysis* (PCA) sebagai langkah *pre-processing* untuk mengurangi dimensi atribut. Atribut baru yang terbentuk dari hasil PCA akan digunakan untuk melakukan proses klasifikasi. Proses klasifikasi dilakukan untuk mengetahui kualitas air yang bersih dengan menggunakan metode *Logistic Regression* dan *K-Nearest Neighbor* (kNN). Tahap akhir yang dilakukan pada penelitian ini yaitu kinerja hasil klasifikasi dari kedua metode tersebut kemudian akan diperbandingkan dan dianalisis untuk mengetahui bagaimana pengaruh metode reduksi dimensi *Principal Component Analysis* (PCA) terhadap proses klasifikasi kualitas air menggunakan metode *Logistic Regression* dan kNN.

2. METODOLOGI PENELITIAN

Penelitian ini dilaksanakan untuk klasifikasi kualitas air layak minum. Data yang digunakan diperoleh dari data *water quality* di kaggle. Data yang didapatkan mula-mula akan dilakukan proses *data cleaning*. *Data cleaning* merupakan teknik untuk menangani data yang tidak lengkap (*missing value),* dengan beberapa cara, seperti membuang duplikasi data, memeriksa data yang inkonsisten, mengisi atau menghapus data kosong, dan lain sebagainya (Jasmir, 2016).

Pada penelitian ini *data cleaning* dilakukan dengan cara menghapus baris data yang kosong. Setelah itu, reduksi dimensi menggunakan metode PCA akan dilakukan untuk mendapatkan fitu-fitur baru hasil reduksi dimensi. Fitur hasil reduksi dimensi ini kemudian akan digunakan sebagai input untuk proses klasifikasi kualitas air layak minum menggunakan metode kNN dan *Logistic Regression*. Tahapan pada penelitian dapat dilihat pada gambar 2.1.



Gambar 1 Tahapan Penelitian

A. Data Penelitian

Penelitian ini menggunakan data sekunder, yang diperoleh dari website yang menyediakan dataset yaitu (Kaggle, 2021) tentang data kualitas air di lingkungan perkotaan, dengan jumlah data 8000 dan 21 atribut. Pada Tabel 1 disajikan dataset kualitas air yang merupakan kandungan dari air dengan keterangan yaitu A = *Alumunium*; B =

Ammonia; C = *Arsenic*; D = *Barium*, E = *Cadmium*, F = *Clhoraminium*, G = *Chromium*, H = *Copper*, I = *Flouride*, J = *Bacteria*, K = *Viruses*, L = *Lead*, M = *Nitrates*, N = *Nitrites*, O = *Mercury*, P = *Perchlorate*, Q = Radium, R = *Selenium*, S = *Silver*, T = *Uranium*, dan U = *Is_safe*.

B. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah suatu metode yang biasa digunakan sebagai alat untuk mereduksi dimensi data, menjadi bentuk yang berada pada bidang nilai yang berbeda (Raysyah, Arinal and Mulyana, 2021). PCA digunakan untuk mengekstraksi struktur dari suatu set data dengan dimensi yang cukup banyak, sebagaimana yang telah dijelaskan oleh (Ilmaniati and Putro, 2019) bahwa metode *Principal Component Analysis* (PCA) lebih tepat digunakan jika tujuan penelitian adalah untuk meringkas data dengan jumlah variabel yang lebih kecil. Analisis ini juga bisa digunakan saat ingin menguji apakah variabel yang sedang diteliti saling bergantung atau justru tidak terkait sama sekali. Algoritma PCA secara umum sebagai berikut (Muhtadi, 2017) :

1. Hitung rata-rata atau *mean* dari data.
2. Hitung Matriks Kovarian, menggunakan Persamaan 1.

$$Cov(xy) = \frac{\sum(xi - \bar{x})(yi - \bar{y})}{n - 1} \quad (1)$$

dimana x dan y adalah dua variabel yang berbeda yang ada pada data, sedangkan $\bar{x}$ dan $\bar{y}$ adalah rata-rata dari variabel x dan y, serta n adalah jumlah data.

3. Hitung nilai eigen dengan Persamaan 2.

$$(A - \lambda I) = 0 \quad (2)$$

dimana A adalah sebuah matriks, dan λ adalah nilai eigen dari A.

4. Hitung nilai vektor eigen dengan Persamaan 3.

$$[A - \lambda I][X] = 0 \quad (3)$$

5. Tentukan variabel baru dengan cara mengalikan variabel asli dengan matriks vektor eigen.

C. K-Nearest Neighbor (KNN)

Metode *K-Nearest Neighbor* (kNN) merupakan salah satu metode dasar yang berasal dari kelompok *instance-based learning* (Kripsiandita, Arifianto and A'yun, 2021). Algoritma kNN melakukan klasifikasi data baru berdasarkan jarak pembelajaran yang paling dekat dengan objek tersebut atau biasa disebut tetangga terdekat (Yustanti, 2012). Algoritma kNN menggunakan pembelajaran *supervised learning* dimana hasil dari data yang diuji diklasifikasikan berdasarkan keanggotaan terdekat yang terbanyak dari data uji (Novita, Harsani and Qur'ania, 2018). Jumlah tetangga paling dekat disebut K. Nilai K dapat ditentukan dengan ketentuan n+1 dimana n adalah jumlah label, atau menggunakan metode ganjil dan genap, jika nilai n ganjil maka nilai k genap, dan sebaliknya. Langkah yang dilakukan untuk menentukan jarak terdekat yaitu, terlebih dahulu data dibagi menjadi data latih dan data uji, setelah diperoleh data latih dan data uji kemudian di hitung jarak masing-masing data uji terhadap data latih (Setianto, Kusrini and Henderi, 2019). Perhitungan jarak dapat menggunakan *Euclidean Distance* dan rumusnya dilihat pada Persamaan 4.

$$d(x_i x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (4)$$

dimana $d(x_i x_j)$ adalah jarak *Euclidean*, $x_i x_j$ data ke i, *record* ke j, a_r adalah data ke-r, n adalah dimensi objek.

D. *Logistic Regression*

p-ISSN : 2502-5724; e-ISSN : 2541-5735

Logistic Regression adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara variabel terikat yang memiliki dua kategori atau lebih dari satu peubah bebas berskala kategori atau kontinyu (Hosmer and Lemeshow, 2000). *Logistic Regression* menunjukkan pengaruh variabel prediktor, baik berupa kontinyu maupun kategorik, teradap variabel respon berupa data kategorik (Purwa, 2019). *Logistic Regression* memiliki beberapa model, salah satunya adalah model regresi logistik biner yang dapat digunakan jika variabel responnya menghasilkan dua kategori bernilai 0 dan 1 (Tampil, Komaliq and Langi, 2017), sehingga mengikuti distribusi bernouli (Agresti, 1990) pada Persamaan 5 :

$$f(y_i) = n\frac{yi}{i}\ (1-\pi_i)^{1-yi} \quad (5)$$

dimana :

π_i = Peluang kejadi ke-i

y_i = peubah acak ke-I yang terdiri dari 0 dan 1

Bentuk model *Logistic Regression* dengan satu variabel prediktor pada Persamaan 6 :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad (6)$$

Transformasi $\pi(x)$ pada persamaan 6 perlu dilakukan untuk mempermudah menaksir parameter regresi, dan menghasilkan bentuk transformasi logit *Logistic Regression* pada Persamaan 7 (Hosmer and Lemeshow, 2000).

$$g(x) = ln\left[\frac{n(x)}{1-n(x)}\right] = \beta_0 + \beta_1 x \quad (7)$$

E. Pengujian

Pengujian dilakukan untuk mengukur model klasifikasi yang dihasilkan. Pengujian pada penelitian ini dilakukan dengan mencari nilai akurasi, presisi dan *recall*. Akurasi yaitu mencari tingkat kedekatan antara nilai aktual dengan nilai prediksi seperti pada Persamaan 8. Sedangkan nilai presisi diukur dari tingkat ketepatan informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem seperti yang terlihat pada Persamaan 9. Nilai *recall* menggambarkan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi (Persamaan 10).

$$Akurasi = \frac{tp + tn}{tp + fn + fp + tn} \quad (8)$$

$$Presisi = \frac{tp}{tp + fp} \quad (9)$$

$$Recall = \frac{tp}{tp + fn} \quad (10)$$

dimana :

a. TP *(True Positif)* : banyaknya data di kelas aktualnya positif dan kelas prediksi positif
b. FN *(False Negative)* : banyaknya data di kelas aktualnya positif dan kelas prediksi negative
c. FP *(False Positive)* : banyaknya data di kelas aktualnya negaitf dan kelas prediksi positif
d. TN *(True Negative)* : banyaknya data di kelas aktualnya negatif dan kelas prediksi Negatif

3. HASIL DAN PEMBAHASAN

3.1 *Preprocessing* Data

Dataset yang digunakan berupa data sekunder yang terdiri dari 8000 baris data dan 21 fitur, seperti yang ditunjukkan pada Tabel 1. Langkah *preprocessing* harus dilakukan terlebih dahulu sebelum melakukan proses klasifikasi. Tahapan *preprocessing* yang dilakukan pada penelitian ini berupa *data cleaning* dan reduksi dimensi.

Tahap *data cleaning* yaitu menangani *missing value*. *Data cleaning* yang dilakukan pada dataset kualitas air berupa penghapusan baris data, karena dari 8000 baris data, terdapat 3 baris yang mengandung *missing value*. Hasil dari *data cleaning* berupa baris data yang awalnya 8000 menjadi 7997 baris data.

Tabel 1 Dataset Kualitas Air

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1.65	9.08	0.04	2.85	0.01	0.35	0.83	0.17	0.05	0.20	0.00	0.05	16.08	1.13	0.01	37.75	6.78	0.08	0.34	0.02	1
2.32	21.16	0.01	3.31	0.00	5.28	0.68	0.66	0.90	0.65	0.65	0.10	2.01	1.93	0.00	32.26	3.21	0.08	0.27	0.05	1
1.01	14.02	0.04	0.58	0.01	4.24	0.53	0.02	0.99	0.05	0.00	0.08	14.16	1.11	0.01	50.28	7.07	0.07	0.44	0.01	0
1.36	11.33	0.04	2.96	0.00	7.23	0.03	1.66	1.08	0.71	0.71	0.02	1.41	1.29	0.00	9.12	1.72	0.02	0.45	0.05	1
0.92	24.33	0.03	0.20	0.01	2.67	0.69	0.57	0.61	0.13	0.00	0.12	6.74	1.11	0.00	16.90	2.41	0.02	0.06	0.02	1
0.94	14.47	0.03	2.88	0.00	0.8	0.43	1.38	0.11	0.67	0.67	0.14	9.75	1.89	0.01	27.17	5.42	0.08	0.19	0.02	1
2.36	5.60	0.01	1.35	0.00	1.28	0.62	1.88	0.33	0.13	0.01	0.02	18.60	1.78	0.01	45.34	2.84	0.10	0.24	0.08	0
3.93	19.87	0.04	0.66	0.00	6.22	0.10	1.86	0.86	0.16	0.01	0.20	13.65	1.81	0.00	53.35	7.24	0.08	0.08	0.07	0

Tahap selanjutnya pada saat *preprocessing* data pada penelitian ini adalah melakukan reduksi dimensi menggunakan metode PCA. Reduksi dimensi dengan metode PCA dilakukan untuk mengetahui fitur-fitur yang terdapat pada dataset saling berhubungan atau tidak. Fitur pada dataset berjumlah 21, dan dibagi menjadi 2 bagian yaitu 1 label, dimana 1 label ini adalah variable is_safe dan 20 variabel lainnya akan menjadi fitur untuk klasifikasi, dimana sebelumnya dilakukan reduksi dimensi terlebih dahulu menggunakan metode PCA.

3.2 Pemrosesan Data

Pemrosesan data dilakukan setelah *preprocessing* data. Tahap pemrosesan data yang dilakukan adalah proses klasifikasi dengan 2 metode yaitu metode kNN dan *Logistic Regression*.

Data yang sudah selesai di *preprocessing* dibagi kembali menjadi data latih dan data uji sebelum dilakukan klasifikasi. Data latih akan digunakan dalam membuat model kNN dan *Logistic Regression*, sedangkan data uji digunakan untuk menguji performa dari kedua metode.

Metode *Logistic Regression* dibangun dengan random_state=0, n_jobs=20, dan max_iter 120. Metode kNN yang dibangun dengan menggunakan jumlah tetangga terdekat yaitu dimulai dari n+1, lalu selanjutnya angka ganjil, karena jumlah label genap, sehingga ditentukan jumlah tetangga k=3, k=5, k=7, k=9, k=11, k=13. Diharapkan dengan eksperimen menggunakan poin-poin pada masing-masing metode tersebut bisa didapatkan hasil terbaik.

3.3 Pengujian

Penelitian ini menggunakan skema pengujian dengan proporsi sampel untuk uji untuk data latih dan data uji sebesar 80%:20%. Pengujian dilakukan dengan 2 tahap yaitu pengujian sebelum penerapan metode PCA dan pengujian setelah Penerapan metode PCA untuk mengetahui pengaruh PCA pada kedua metode klasifikasi.

Pengujian untuk metode kNN yang dilakukan sebelum penerapan metode PCA didapatkan hasil akurasi, presisi dan *recall* seperti ditunjukkan pada tabel 2. Hasil pengujian untuk metode kNN dengan penerapa PCA ditunjukkan pada Tabel 3 sampai Tabel 6.

Tabel 2 Hasil Pengujian kNN Tanpa PCA

Nilai k	Akurasi	Presisi	Recall
3	88.0%	87.0%	88.0%

p-ISSN : 2502-5724; e-ISSN : 2541-5735

5	90.0%	89.0%	90.0%
7	90.6%	90.0%	91.0%
9	90.8%	90.0%	91.0%
11	90.6%	90.0%	91.0%
13	90.3%	90.0%	90.0%

Tabel 2 menunjukkan hasil dari pengujian metode kNN tanpa PCA, didapatkan hasil akurasi tertinggi 90.6%, presisi tertinggi 90.0%, dan *recall* tertinggi 91.0% yang terdapat pada k=9.

Tabel 3 Hasil Pengujian kNN+PCA n=2

Nilai k	2-Dimensi		
	Akurasi	Presisi	Recall
3	86.0%	84.0%	86.0%
5	86.6%	83.0%	87.0%
7	86.4%	81.0%	86.0%
9	87.0%	82.0%	87.0%
11	87.4%	83.0%	87.0%
13	87.0%	82.0%	87.0%

Tabel 3 menunjukkan hasil dari pengujian metode kNN dengan PCA yang berdimensi 2, didapatkan hasil akurasi tertinggi 87.4%, presisi tertinggi 84.0%, dan *recall* tertinggi 87.0%.

Tabel 4 Hasil Pengujian kNN+PCA n=3

Nilai k	2-Dimensi		
	Akurasi	Presisi	Recall
3	87.2%	85.0%	87.0%
5	88.3%	86.0%	88.0%
7	87.8%	85.0%	88.0%
9	88.4%	86.0%	88.0%
11	88.3%	86.0%	88.0%
13	88.4%	86.0%	88.0%

Tabel 4 menunjukkan hasil dari pengujian metode kNN dengan PCA yang berdimensi 3, didapatkan hasil akurasi tertinggi 88.4%, presisi tertinggi 86.0%, dan *recall* tertinggi 88.0% yang terdapat pada k=9 dan k=13.

Tabel 5 Hasil Pengujian kNN+PCA n=4

Nilai k	2-Dimensi		
	Akurasi	Presisi	Recall
3	87.6%	85.0%	88.0%
5	87.8%	85.0%	88.0%
7	87.8%	85.0%	88.0%
9	87.9%	85.0%	88.0%
11	88.5%	86.0%	89.0%
13	88.3%	86.0%	88.0%

Tabel 5 menunjukkan hasil dari pengujian metode kNN dengan PCA yang berdimensi 4, didapatkan hasil akurasi tertinggi 88.5%, presisi tertinggi 86.0%, dan *recall* tertinggi 89.0% yang terdapat pada k=11.

Tabel 6 Hasil Pengujian kNN+PCA n=5

Nilai k	2-Dimensi		
	Akurasi	Presisi	Recall
3	86.3%	84.0%	86.0%
5	88.5%	86.0%	89.0%
7	87.6%	85.0%	88.0%
9	88.0%	85.0%	88.0%
11	87.8%	85.0%	88.0%
13	88.0%	86.0%	88.0%

Tabel 6 menunjukkan hasil dari pengujian metode kNN dengan PCA yang berdimensi 5, didapatkan hasil akurasi tertinggi 88.5%, presisi tertinggi 86.0%, dan *recall* tertinggi 89.0% yang terdapat pada k=5.

Pengujian untuk metode *Logistic Regression* yang dilakukan sebelum penerapan metode PCA didapatkan hasil akurasi 90.0%, presisi 89.0%, dan *recall* 90.0%, seperti ditunjukkan pada Tabel 7.

Tabel 7 Hasil Pengujian *Logistic Regression* Tanpa PCA

Akurasi	Presisi	Recall
90.0%	89.0%	90.0%

Pengujian untuk metode *Logistic Regression* dilakukan juga dengan penerapan metode PCA didapatkan hasil

akurasi, presisi, dan *recall*, seperti ditunjukkan pada Tabel 8.

Tabel 8 Hasil Pengujian *Logistic Regression*+PCA Dengan n_dimensi

Dimensi	Akurasi	Presisi	Recall
2	87.5%	77.0%	88.0%
3	87.6%	84.0%	88.0%
4	87.5%	84.0%	88.0%
5	87.9%	84.0%	88.0%

Tabel 8 menunjukkan hasil dari pengujian metode *Logistic Regression* dengan PCA yang berdimensi n=2, n=3, n=4, dan n=5. Hasil akurasi tertinggi 87.9%, presisi tertinggi 84.0%, dan *recall* tertinggi 88.0% yang terdapat pada n=5.

3.4 Evaluasi dan Analisis

Perbandingan performa klasifikasi model terbaik dari metode kNN dan *Logistic Regression* baik sebelum penerapan PCA maupun setelah penerapan PCA disajikan pada Tabel 9. Kedua metode memiliki nilai akurasi yang tidak jauh berbeda.

Tabel 9 Perbandingan Performa Kedua Metode

	kNN			*Logistic Regression*		
Dimensi	Akurasi	Presisi	Recall	Akurasi	Presisi	Recall
2	87.4%	83.0%	87.0%	87.5%	77.0%	88.0%
3	88.4%	86.0%	88.0%	87.6%	84.0%	88.0%
4	88.5%	86.0%	89.0%	87.5%	84.0%	88.0%
5	88.5%	86.0%	89.0%	87.9%	84.0%	88.0%
Tanpa PCA	90.8%	90.0%	91.0%	90.0%	89.0%	90.0%

Metode kNN tanpa PCA memiliki performa yang lebih baik dibanding dengan penerapan PCA dimana tanpa PCA didapatkan akurasi tertinggi 90.8% pada k=9, sedangkan dengan penerapan PCA akurasi tertinggi yaitu 88.5% pada n=5 dan k=5.

Metode *Logistic Regression* tanpa PCA juga memiliki performa yang lebih baik dibanding dengan penerapan PCA dimana tanpa PCA didapatkan akurasi tertinggi 90.0%, sedangkan dengan penerapan PCA akurasi tertinggi yaitu 87.9% pada n=5.

Hasil perbandingan secara keseluruhan dapat disimpulkan yaitu pada klasifikasi kualitas air, dengan menerapkan metode reduksi dimensi PCA dapat menurunkan tingkat performa metode klasifikasi kNN dan *Logistic Regression*. Hasil penelitian juga menunjukkan bahwa performa metode kNN lebih baik dari metode *Logistic Regression*, baik sebelum maupun sesudah penerapan PCA. Berdasarkan hasil tersebut, maka metode klasifikasi terbaik secara keseluruhan adalah metode kNN tanpa penerapan PCA yaitu dengan akurasi 90.8% pada jumlah k=9.

4. KESIMPULAN

Air bersih dan sanitasi layak adalah kebutuhan dasar manusia yang termasuk salah satu poin dalam tujuan pembangunan berkelanjutan (*sustainable development goals/SDGs*) pada sektor lingkungan hidup. Tingkat aktivitas manusia yang semakin tinggi mengakibatkan tingkat pencemaran juga semakin tinggi, sehingga akan menyebabkan turunnya kualitas air dan akan mempengaruhi akses ke air minum yang aman untuk dikonsumsi oleh manusia.

Penurunan kualitas air menyebabkan urgensi kebutuhan untuk mengawasi, menilai, dan mengklasifikasi kualitas air yang layak untuk konsumsi. Proses klasifikasi kualitas air dapat dilakukan dengan metode kNN dan *Logistic Regression* serta menggunakan metode *Principal Component Analysis* (PCA) sebagai langkah *pre-processing* untuk mengurangi dimensi data.

Hasil penelitian secara keseluruhan yaitu tingkat performa dari metode kNN dan *Logistic Regression* menurun saat menerapkan metode reduksi dimensi PCA. Metode kNN tanpa PCA memiliki akurasi tertinggi 90.8% pada k=9, sedangkan

p-ISSN : 2502-5724; e-ISSN : 2541-5735

metode kNN dengan PCA memiliki akurasi 88.5% pda n=5 dan k=5. Metode *Logistic Regression* tanpa PCA memiliki akurasi tertinggi 90.0%, sedangkan dengan PCA memiliki akurasi 87.9 pada n=5.

DAFTAR PUSTAKA

Agresti, A., 1990. *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Hosmer, D.W. and Lemeshow, S., 2000. *Applied Logistic Regression*. 2nd ed. United States of America: John Wiley & Sons, Inc.

Ilmaniati, A. and Putro, B.E., 2019. Analisis komponen utama faktor-faktor pendahulu (antecendents) berbagi pengetahuan pada usaha mikro, kecil, dan menengah (UMKM) di Indonesia. *Jurnal Teknologi*, [online] 11(1), pp.67–78. Available at: <https://jurnal.umj.ac.id/index.php/jurtek/article/view/2652>.

Jasmir, 2016. Implementasi Teknik Data Cleaning dan Teknik Roughset pada Data Tidak Lengkap dalam Data Mining. *Seminar Nasional APTIKOM (SEMNASTIKOM)*, pp.99–106.

Kaggle, 2021. *Water Quality Dataset*.

Kripsiandita, Y., Arifianto, D. and A'yun, Q., 2021. Deteksi Gangguan Autis Pada Anak Menggunakan Metode Modified K-Nearst Neighbor. *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, 6(1), pp.31–37. https://doi.org/10.32528/justindo.v6i1.4357.

Kustanto, A., 2020. Water quality in Indonesia: The role of socioeconomic indicators. *Jurnal Ekonomi Pembangunan*, 18(1), pp.47–62. https://doi.org/10.29259/jep.v18i1.11509.

Muhtadi, 2017. Penerapan Principal COmponent Analysis (PCA) Dalam Algoritma K-Means Untuk Menentukan Centroid Pada Clustering. *Journal of Mathematic Teaching*, 1(1), pp.121–142.

Novita, S., Harsani, P. and Qur'ania, A., 2018. Penerapan K-Nearest Neighbor (KNN) untuk Klasifikasi Anggrek Berdasarkan Karakter Morfologi Daun dan Bunga. *Komputasi*, 15(1), pp.118–125.

Pritalia, G.L., 2022. Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum. *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, 2(1), pp.43–55. https://doi.org/10.24002/konstelasi.v2i1.5630.

Purwa, T., 2019. Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017). *Jurnal Matematika, Statistika dan Komputasi*, 16(1), p.58. https://doi.org/10.20956/jmsk.v16i1.6494.

Rahman, M.A., Hidayat, N. and Supianto, A.A., 2018. Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 12, Desember 2018, hlm. 6346-6353 e-ISSN:*, 2(12), pp.925–928.

Raysyah, S., Arinal, V. and Mulyana, D.I., 2021. Klasifikasi Tingkat Kematangan Buah Kopi Berdasarkan Deteksi Warna Menggunakan Metode Knn Dan Pca. *JSiI (Jurnal Sistem Informasi)*, 8(2), pp.88–95. https://doi.org/10.30656/jsii.v8i2.3638.

SDGS, S.N., 2022. *Air Bersih dan Sanitasi Layak*. [online] SDGS Bappenas. Available at: <https://sdgs.bappenas.go.id/tujuan-6/> [Accessed 27 June 2022].

Setianto, Y.A., Kusrini, K. and Henderi, H., 2019. Penerapan Algoritma K-Nearest Neighbour Dalam Menentukan Pembinaan Koperasi Kabupaten Kotawaringin Timur. *Creative Information Technology Journal*, 5(3), p.232. https://doi.org/10.24076/citec.2018v5i3.179.

Tampil, Y., Komaliq, H. and Langi, Y., 2017. Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado. *d'CARTESIAN*, 6(2), p.56. https://doi.org/10.35799/dc.6.2.2017.17023.

Yustanti, W., 2012. Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika statistika dan komputasi*, 9(1), pp.57–68.