Jurnal TIKomSiN, Vol. 6, No. 1, April 2018
ISSN Cetak : 2338-4018
ISSN Online : 2620-7532

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR UNTUK IDENTIFIKASI KUALITAS AIR (STUDI KASUS: PDAM KOTA SURAKARTA)

Rio Adi Arnomo[1)]; Wawan Laksito Yuly Saptomo [2)]; Paulus Harsadi [3)]
[1) 2) 3)] Program Studi Teknik Informatika, STMIK Sinar Nusantara
[1)]akunrio09@gmail.com; [2)]wlaksito@sinus.ac.id; [3)]paulusharsadi@sinus.ac.id

ABSTRACT

Water quality in urban areas in Surakarta has decreased nowadays. The increase of industrial development, its poor sewage treatment, and some other factors cause this urban problem. The result of water quality monitoring system with K-Nearest Neighbor algorithm on this research certainly will help the laborers' duty of PDAM (Local Government Owned Water Utilities) in analyzing water quality. For the consideration of majority output in this method, the system works by taking the nearest distance to the assigned number of K. The training data for this research was taken in March 2016 form the report of water monitoring result in PDAM'S laboratory of Surakarta. The identification result is divided into eligible (MS) and ineligible (TMS). The testing data result is applied in algorithm performance testing with confusion matrix having accuracy level 82,5%.

Keywords : Water Quality, K-Nearest Neighbor, Confusion Matrix

I. PENDAHULUAN

Air merupakan kebutuhan penting bagi masyarakat perkotaan pada khususnya. Kebutuhan air bersih tiap tahun mengalami peningkatan sedangkan ketersediaan air bersih semakin terbatas, dikarenakan sempitnya daerah resapan, banyaknya pembangunan yang tidak memperhatikan keseimbangan alam, eksploitasi sumber air baku yang tidak terjaga kelestarian sumber air. Salah satu masalah pokok yang dihadapi adalah kurang tersedianya sumber air bersih.

Berdasarkan Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/MENKES/PER/IV/2010 tentang Persyaratan Kualitas Air Minum terdapat pengertian mengenai Air Minum yaitu air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum. Kualitas air yang layak konsumsi bagi masyarakat, perlu adanya identifikasi dini terhadap produk air dari sumber air baku serta faktor-faktor yang mempengaruhinya. Tujuan penelitian ini untuk melakukan *early warning* untuk identifikasi kualitas air dengan mengklasifikasikan data menggunakan metode algoritma *K-Nearest Neighbor (K-NN)*. Kontribusi implementasi algoritma K-NN untuk lingkungan sendiri sudah cukup banyak antara lain untuk peramalan temperatur air[1], peramalan banjir [2] bahkan dalam perubahan iklim [3]. Algoritma ini termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learning*. KNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. Algoritma K-Nearest Neighbor adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut[4]. K-NN sendiri merupakan teknik yang sudah umum digunakan dalam berbagai bidang penelitian salah satunya *text categorization* [5][6][7] bahkan dalam *object recognition* [8].

II. TINJAUAN PUSTAKA

2.1. Sistem Penunjang Keputusan

Sistem Pendukung Keputusan mempunyai komponen-komponen adalah sebagai berikut [9]:

a. Data Management
   Termasuk database yang mengandung data yang relevan untuk berbagai situasi dan diatur oleh software.
b. Model Management
   Melibatkan model finansial, statistikal, management science, atau berbagai model kualitatif lainya, sehingga dapat memberikan ke sistem suatu kemampuan analitis dan manajemen software yang dibutuhkan.
c. Communication / Interface
   User dapat berkomunikasi dan memberikan perintah pada DSS melalui subsistem ini.

ISSN Cetak : 2338-4018
ISSN Online : 2620-7532

d. Knowledge Management
Subsistem optional ini dapat mendukung subsistem lain atau bertindak sebagai komponen yang berdiri sendiri.

2.2. Kualitas Air

Air merupakan salah satu sumber daya alam yang memiliki fungsi sangat penting bagi kehidupan dan perikehidupan manusia, serta untuk memajukan kesejahteraan umum, sehingga merupakan modal dasar dan faktor utama pembangunan. Air yang dikonsumsi oleh masyarakat harus memiliki standar air bersih yang telah ditetapkan oleh KEMENKES no.492/Menkes/Per/IV/2010 dalam peraturan yang diterbitkan pada tahun 2010.

Tabel 1. Kadar yang diperbolehkan

No	Jenis Parameter	Satuan	Kadar yang ditentukan
1	Parameter fisik		
	a) Bau		Tak berbau
	b) Rasa		Tak berrasa
	c) Warna	^{o}C	
	d) Suhu	TCU	Suhu udara $\pm 3^{o}$
	e) Kekeruhan	NTU	5
2	Parameter kimia		
	a) Sisa Clor	mg/l	0,2 – 0,5
	b) Besi	mg/l	0,3
	c) Mangan	mg/l	0,4
	d) Kesadahan	mg/l	500
	e) Nitrit	mg/l	3
	f) Amonium	mg/l	1,5
	g) Kalium permanganat	mg/l	
	h) Ph		6,5 – 8,5
	i) Klorida	mg/l	250
	j) sulfat	mg/l	250

III. METODE PENELITIAN

3.1 Data Primer

Data yang dibutuhkan dalam penelitian ini adalah hasil wawancara dan observasi mengenai uji kualitas air di laboratorium PDAM Kota Surakarta.

3.2 Data Sekunder

Data sekunder umumnya berupa bukti, catatan atau laporan historis yang tersusun dalam arsip. Data yang dibutuhkan dalam penelitian ini adalah laporan uji lab kualitas air bulan Maret dan April 2016 di PDAM kota Surakarta sebagai data training sebanyak 111 data.

3.3 Metode K-Nearest Neighbor

Algoritma K-NN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan K-NN-nya. *Training sample* diproyeksikan ke ruang berdimensi banyak dimana tiap dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi *training sample.* Sebuah titik pada ruang ini ditandai kelas *c* jika kelas *c* merupakan klasifikasi yang paling banyak ditemui pada *k* buah tetangga terdekat dari titik tersebut [10]. Untuk mendefinisikan jarak antara dua titik yaitu titik pada data training (x) dan titik pada data testing (y) maka digunakan rumus Euclidean, seperti yang ditunjukkan pada persamaan

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

Dengan D adalah jarak antara titik pada data training x dan titik data testing yang akan diklasifikasi, dimana x=x1,x2,...,xi dan y1,y2,...,yi dan I mempresentasikan nilai atribut serta n merupakan dimensi atribut. Langkah-langkah untuk menghitung metode *Algoritma K-nearest Neighbor* [4]:

a. Menentukan Parameter K (Jumlah tetangga paling dekat).
b. Menghitung kuadrat jarak *Euclid* (*query instance*) masing-masing objek terhadap data sampel yang diberikan
c. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak *Euclid* terkecil.
d. Mengumpulkan kategori Y (*Klasifikasi Nearest Neighbor*).
e. Dengan menggunakan kategori *Nearest Neighbor* yang paling mayoritas maka dapat diprediksi nilai *query instance* yang telah dihitung.

3.4 Pengujian Confusion matrix

Metode ini menggunakan tabel matriks seperti padaTabel 1 jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif [11].

Tabel 2 Model Confusion Matrix

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	true positives	false negatives
-	false positives	true negatives

True positives adalah jumlah record positif yang diklasifikasikan sebagai positif, false positives adalah jumlah record negatif yang diklasifikasikan sebagai positif, false negatives adalah jumlah record positif yang diklasifikasikan sebagai negatif, true negatives adalah jumlah record negatif yang diklasifikasikan sebagai negative, kemudian

ISSN Cetak : 2338-4018
ISSN Online : 2620-7532

masukkan data uji. Setelah data uji dimasukkan ke dalam confusion matrix, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah sensitivity (recall) specificity, precision dan accuracy. Sensitivity digunakan untuk membandingkan jumlah TP terhadap jumlah record yang positif sedangkan specificity adalah perbandingan jumlah TN terhadap jumlah record yang negatif. Untuk menghitung digunakan persamaan di bawah ini [12]:

$$sensitivity = \frac{TP}{P} \quad (2)$$

$$specificity = \frac{TN}{N} \quad (3)$$

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$accuracy = sensitivity\frac{P}{(P+N)} + specificity\frac{N}{(P+N)} \quad (5)$$

Keterangan:

TP = jumlah true positives
TN = jumlah true negatives
P = jumlah record positif
N = jumlah tupel negatif
FP = jumlah false positives

IV. HASIL DAN PEMBAHASAN

4.1. Pembahasan

Penelitian ini menggunakan 111 *record* hasil analisa kualitas air baik yang Memenuhi Syarat ataupun Tidak Memenuhi Syarat diambil dari laporan analisa kualitas air PDAM Surakarta. Berikut pada Tabel 3 menjelaskan variabel respon yang digunakan dalam penelitian ini.

Tabel 3. Variabel respon

ASPEK	VARIABEL
	Y : Status Kualitas Hasil Pemeriksaan Sample Air (1 = MS (Memenuhi Syarat), 2 = TMS (Tidak Memenuhi Syarat)
Parameter fisik	X_1 : Bau
	X_2 : Rasa
	X_3 : Suhu
	X_4 : Warna
	X_5 : Kekeruhan
Parameter Kimia	X_6 : Sisa Clor
	X_7 : pH
	X_8 : Kesadahan
	X_9 : Besi
	X_{10} : Mangan
	X_{11} : Amonium
	X_{12} : Nitrit
	X_{13} : kalium permanganat
	X_{14} : Klorida
	X_{15} : Sulfat

Data yang nilainya bersifat nominal akan diubah menjadi data numerik. Pada Tabel 3 data yang masih bersifat nominal adalah bau dan rasa. Tabel 4 merupakan perubahan data nominal ke numerik.

Tabel 4. Nilai atribut bau dan rasa

Atribut	Nilai
Berbau	1
Tak berbau	2
Berrasa	1
Tak berrasa	2

Perhitungan dilakukan dengan 15 variabel data yaitu data fisik dan numerik. Data testing akan dihitung jarak terdekat dengan masing-masing data training. Misalkan sebuah data sampel kualitas air yang berlokasi di Toko Djoyo apakah memenuhi syarat atau tidak memenuhi syarat maka akan dihitung jarak terdekat dengan data training yang sudah ada.

Contoh perhitungan dengan algoritma *K-Nearest Neighbor:*

a. Menentukan parameter K, misal K = 7
b. Menghitung kuadrat jarak euclidean (euclidean distance) masing-masing obyek terhadap data sampel yang diberikan

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

d(1,1) =

$= ((0)^2 + (0)^2 + (4)^2 + (0)^2 + (-0,04)^2 + (0)^2 + (0,5)^2 + (12)^2 + (0,07)^2 + (-0,79)^2 + (0,13)^2 + (0)^2 + (0,01)^2 + (26)^2 + (1)^2)^{1/2}$

$= (0 + 0 + 16 + 0 + 0,0016 + 0 + 0,25 + 144 + 0,0049 + 0,624 + 0,0169 + 0 + 0,01 + 676 + 1)^{1/2}$

$= (837,8976)^{1/2}$

=28,95

(dan seterusnya sampai d(1,71)).

ISSN Cetak : 2338-4018
ISSN Online : 2620-7532

Tabel 5. Data Training

No. Lab	Lokasi	Parameter Fisika					Parameter Kimia										Keterangan
		Bau	Rasa	Suhu	Warna	Kekeruhan	Sisa Clor	pH	kesadahan	Fe	Mn	NH_4^+	NO_2^-	$KMnO_4$	Cl^-	SO_4^{2-}	
190	Manahan 1	1	1	30	0	0,41	0	7,3	204	0,42	0,11	0,84	0,001	0,32	50	11,5	TMS
191	Manahan 2	1	1	30	0	0	0	7,3	146	0,03	0,09	2	0	0	43,3	8,17	MS
198	Bak Manahan	1	1	30	10	2,8	0	7,3	146	0,36	0,47	0,91	0,004	0,32	49	6,45	MS
192	Tirtonadi	1	1	30	0	0	0	7,3	194	0,09	0	0,81	0	0	37	11,3	TMS
193	Banjarsari	1	1	30	0	0	0	7,3	166	0,02	0,02	0,95	0,001	0	49	3,9	MS
199	Bak Banjarsari	1	1	30,5	0	0	0	7,3	144	0	0,02	0,15	0,002	0	50,96	3,9	MS
194	Jebres 1	1	1	30	0	0	0	7,1	164	0,02	0,29	0,98	0,008	0,02	88	7,1	MS
195	Jebres 2	1	1	29	0	0	0	7,1	168	0	0,22	0,41	0,017	0	38,5	11,3	MS
202	IPA Jebres	1	1	30	0	3,28	0	7,1	112	0,1	0,09	0,25	0,011	2,45	29,81	10,97	MS
200	Bak Jebres	1	1	30	0	0	0,2	7,9	124	0,02	0,25	0,68	0,009	1,54	64,9	9,4	MS
196	Pedaringan	1	1	30	0	0	0	7,1	112	0	0	0,15	0,002	0,62	31,25	11,29	MS
197	Jurug 2	1	1	30	0	0	0	7,8	90	0	0	0,85	0	0	14,74	6,45	MS
204	IPA Jurug	1	1	30	0	0	0	7,1	114	0,1	0,01	0,25	0,011	2,45	29,8	11	MS
205	Mess Bengawan Solo,	1	1	29	0	0	0	7,1	124	0,05	0	0,11	0	0	19,7	7,1	MS
206	Setyo Budiarto, Jl.AR H	1	1	30	0	0	0	7,1	144	0,02	0,16	0,17	0	0,002	41,8	97	MS
207	Partiyem Heru, Tegal K	1	1	29	0	0	0	7,1	84	0,02	0,02	0,21	0,002	0,62	28,3	10,7	MS
208	Siti Rochayani, Gulon	1	1	29	6	0	0	7,3	84	0	0,02	0,13	0,002	1,23	32,21	10	MS
209	Masjid Al Muharijin, J	1	1	30	6	0	0	7,5	80	0,03	0,09	0,14	0	0,62	26,4	8,3	MS
210	Tiurlan Petoran RT 3/4	1	1	30	0	0	0	7,1	128	0,02	0,2	0,25	0,002	0	38,5	10,8	MS
211	Kantor PAC PDIP, Ganj	1	1	30	0	0	0	7,1	124	0,03	0,11	0,14	0,006	1,23	26,4	9,4	MS
219	Banyuanyar	1	1	30	0	0	0	7,3	134	0,1	0,36	0,54	0	0	36,41	4,52	MS
220	Kadipiro 1	2	1	30	0	0	0	7,6	152	0,05	0	0,24	0,02	0	11,17	4	MS
221	Kadipiro 2	1	1	30	0	0	0	7,3	136	0	0,29	0,3	0,024	0	25,2	8,2	MS
222	Kadipiro 3	1	1	30	0	0	0	6,9	148	0	0,04	0,12	0,012	0,62	26,7	5,8	MS

Tabel 6. Data Testing

No. Lab	Lokasi	Parameter Fisika					Parameter Kimia										Distance
		Bau	Rasa	Suhu	Warna	Kekeruhan	Sisa Clor	pH	kesadahan	Fe	Mn	NH_4^+	NO_2^-	$KMnO_4$	Cl^-	SO_4^{2-}	
190	Toko Djoyo, JL. Cikaran	1	1	26	0	0,45	0	6,8	192	0,35	0,9	0,71	0,001	0,31	24	10,5	?

4.2 Pengujian

Pengujian algoritma dengan *confusion matrik* digunakan untuk menghitung nilai precision, recall dan accuracy. Perhitungan kedekatan kasus lama pada data *training* dengan kasus baru pada data *testing*, diketahui dari 40 data, 33 data kelas MS diprediksi sesuai, 0 data diklasifikasikan kelas TMS, 7 data yang diprediksi kelas TMS ternyata masuk kelas MS, dan 0 data yang diprediksi MS ternyata masuk kelas TMS. Berikut tabel 5 model *confusion matrix* untuk algoritma KNN.

Tabel 5. Model *Confusion Matriks* untuk algoritma KNN

Nilai Prediksi	Nilai Sebenarnya	
	MS	TMS
Prediksi MS	33	0
Prediksi TMS	7	0

Dari Tabel 5 tersebut akan dihitung nilai precision, recall dan accuracy. Berikut perhitungannya:

Precision=$\frac{TP}{TP+FP} = \frac{33}{33+7} = \frac{33}{40} = 0{,}825 = 82{,}5\%$

Recall = $\frac{TP}{TP+FN} = \frac{33}{33+0} = \frac{33}{33} = 1 = 100\%$

Accuracy = $\frac{TP+TN}{TP+FP+TN+FN} = \frac{33+0}{33+7+0+0} = \frac{33}{40} = 0{,}825 = 82{,}5\%$

Sehingga diperoleh nilai precision 82,5%, recall 100%, dan accuracy 82,5%.

V. PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dalam menganalisa dan mengimplentasikan sistem identifikasi kualitas air dengan menggunakan algoritma *K-Nearest Neighbor* yang dilakukan oleh peneliti, maka dapat diambil beberapa kesimpulan, yaitu:

1. Penelitian ini berhasil menerapkan algortma *K-Nearest Neighbor* untuk menghitung dan memberikan hasil klasifikasi terhadap kualitas air.
2. Penelitian ini menggunakan*111 record* data, 71 data digunakan sebagai data *training*, 40 data sebagai data *testing.* Dari data *testing* tersebut 33 data diprediksi benar dan 7 data diprediksi salah dengan perhitungan jumlah K sebanyak 7.

ISSN Cetak : 2338-4018
ISSN Online : 2620-7532

3. Sistem untuk identifikasi kualitas air dengan menggunakan metode KNN hasilnya akurat dengan mencapai tingkat akurasi sebesar 82,5%.

5.2. Saran

Saran penelitian berikutnya adalah untuk menghasilkan akurasi tinggi sebaiknya digabungkan menggunakan algoritma optimasi yaitu GA (Genetik *Algorithm*), PSO (*Particle Swam Optimzation*), atau CE (*Cross-Entropy*).

DAFTAR PUSTAKA

[1] A. St-Hilaire, T. B. M. J. Ouarda, Z. Bargaoui, A. Daigle, and L. Bilodeau, "Daily river water temperature forecast model with a k-nearest neighbour approach," *Hydrol. Process.*, vol. 26, no. 9, pp. 1302–1310, 2012.

[2] K. Liu, Z. Li, C. Yao, J. Chen, K. Zhang, and M. Saifullah, "Coupling the k-nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting," *Int. J. Sediment Res.*, vol. 31, no. 2, pp. 149–158, 2016.

[3] H.-I. Eum, S. P. Simonovic, and Y.-O. Kim, "Climate Change Impact Assessment Using K-Nearest Neighbor Weather Generator: Case Study of the Nakdong River Basin in Korea," *J. Hydrol. Eng.*, vol. 15, no. 10, pp. 772–785, 2010.

[4] R. I. Ndaumanu and M. R. Arief, "Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor," *JISATI*, vol. 1, no. 1, pp. 1–15, 2014.

[5] E. M. Elnahrawy, "Log-Based Chat Room Monitoring Using Text Categorization : A Comparative Study," *IASTED Int. Conf. Inf. Knowl. Shar. (IKS 2002)*, 2002.

[6] G. Toker and Ö. Kırmemiş, "Text Categorization Using k-Nearest Neighbor Classification," *Middle East Tech. Univ.*, 2013.

[7] Y. Liao and V. R. Vemuri, "Using Text Categorization Techniques for Intrusion Detection," *Proc. 11 th USENIX Secur. Symp.*, 2002.

[8] F. Bajramovic, F. Mattern, N. Butko, and J. Denzler, "A comparison of nearest neighbor search algorithms for generic object recognition," *Adv. Concepts Intell. Vis. Syst.*, pp. 1186–1197, 2006.

[9] A. . Rosa and A. Shalahuddin, *Modul Pembelajaran Perangkat Lunak (Terstruktur dan Berorientasi Objek)*. Bandung: Modula, 2011.

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[11] M. Bramer, *Principles of Data Mining*, no. January 2007. London: Springer London, 2007.

[12] K. Michelin, P. Jian, and H. Jiawei, "Data Mining: Concepts and Techniques," vol. 278, pp. 6093–6100, 2003.