

KLASIFIKASI KUALITAS AIR PDAM MENGGUNAKAN ALGORITMA KNN DAN K-MEANS

Nurmahaludin¹, Gunawan Rudi Cahyono²
Politeknik Negeri Banjarmasin¹
Universitas Lambung Mangkurat²
mahaludin@poliban.ac.id¹, m4sgunnana@gmail.com²

ABSTRACT

K-Nearest Neighbor (KNN) algorithm is one of classification methods that the number of nearest neighbors (K) must be determined in advance. Data is grouped based on the distance to the nearest neighbor to then calculate the majority class in the group. In this study KNN will be compared with the K-Means algorithm. The K-Means algorithm is included in the unsupervised classification because the class of data to be grouped is unknown in advance.

The two algorithms are used for classification of water quality contained in PDAM Bandarmasih. Water quality parameters used include microbiology, inorganic chemistry, physical parameters, and chemical parameters. The quality of the water tested is classified as feasible and inappropriate.

Based on the test results obtained better accuracy with KNN is 88% compared to K-Means which is 76%.

Keywords: *K-Means, K-Nearest Neighbor, classification, water quality*

ABSTRAK

Algoritma K-Nearest Neighbor (KNN) merupakan salah satu metode yang terdapat dalam klasifikasi. Data dikelompokkan berdasarkan jarak terhadap tetangga terdekatnya untuk kemudian dihitung kelas mayoritas yang ada dalam kelompok tersebut. Dalam penelitian ini KNN akan dibandingkan dengan algoritma K-Means. Algoritma K-Means termasuk dalam *unsupervised classification* karena sejumlah data yang akan dikelompokkan tidak diketahui kelasnya terlebih dahulu.

Kedua algoritma tersebut digunakan untuk klasifikasi kualitas air yang terdapat pada PDAM Bandarmasih. Parameter kualitas air yang digunakan meliputi mikrobiologi, kimia anorganik, parameter fisik, serta parameter kimia. Kualitas air yang diuji diklasifikasikan menjadi layak dan tidak layak.

Berdasarkan hasil pengujian diperoleh akurasi dengan KNN lebih baik yaitu 88% dibandingkan dengan K-Means yaitu 76%.

Kata Kunci: *K-Means, K-Nearest Neighbor, Klasifikasi, Kualitas Air*

PENDAHULUAN

Dalam teknik klasifikasi, suatu data akan dianalisa menggunakan metode tertentu untuk kemudian dimasukkan ke dalam suatu kelas dari sejumlah kelas yang ada. Dalam K-Nearest Neighbor (KNN), jumlah tetangga terdekat (K) harus ditentukan terlebih dahulu. Data dikelompokkan berdasarkan jarak terhadap tetangga terdekatnya untuk kemudian dihitung kelas mayoritas yang ada dalam kelompok tersebut. KNN digunakan antara lain untuk klasifikasi penyakit jantung menggunakan KNN dan algoritma genetika (Jabbar, 2013). Dalam kaitannya dengan penentuan kualitas air, penelitian yang telah ada yaitu perbandingan antara metode KNN dengan Naïve Bayes dalam klasifikasi kualitas air bersih (Rahman, 2018). Sebagaimana diketahui bahwa algoritma KNN termasuk dalam *supervised classification* yang mana pada proses pelatihan diperlukan informasi kelas terlebih dahulu pada data *training*.

K-Means merupakan salah satu algoritma *clustering* yang banyak digunakan dalam berbagai bidang aplikasi, diantaranya dalam bidang pendidikan yaitu sistem pendukung keputusan pemberian beasiswa (Nuraziq, 2017) dan dalam bidang kesehatan untuk diagnosa penyakit gagal ginjal kronis (Musdalifah, 2014). Selain itu K-Means juga digunakan untuk penentuan lokasi wisata (Masruro, 2016) serta untuk membantu dalam analisis kinerja karyawan (Sarker, 2018). K-Means termasuk. Berbeda dengan KNN, K-Means termasuk dalam *unsupervised classification* dimana data yang akan dikelompokkan tidak diketahui kelasnya terlebih dahulu. Dalam penelitian ini dilakukan perbandingan algoritma K-Nearest Neighbor dengan K-Means untuk klasifikasi kualitas air pada PDAM Bandarmasih.

Algoritma K-Means

Pada algoritma K-Means, jumlah kluster K sudah ditentukan lebih dahulu. Setiap kluster memiliki titik pusat dan anggota-anggota dari satu cluster dipilih berdasarkan jarak dari titik pusat kluster terdekat.

Algoritma K-means adalah sebagai berikut (Prasetyo, 2012):

1. Tentukan jumlah kluster (K) yang ingin dibentuk
2. Alokasikan data ke dalam kluster secara random atau dapat juga dilakukan dengan membangkitkan titik pusat kluster (sentroid) awal secara random
3. Hitung jarak setiap data ke masing-masing pusat kluster menggunakan rumus jarak *Euclidean*
4. Kelompokkan tiap data berdasarkan jarak terdekat dengan sentroidnya
5. Tentukan pusat kluster baru dengan cara menghitung nilai rata-rata dari data-data yang ada pada sentroid yang sama
6. Kembali ke langkah 3 jika posisi sentroid baru dengan sentroid lama tidak sama atau apabila masih ada data yang berpindah kluster.

Lokasi sentroid tiap kluster yang diambil dari rata-rata (*mean*) semua nilai data pada tiap fitur harus dihitung kembali. Jika M adalah jumlah data dalam suatu kluster dan i adalah fitur ke-i dalam sebuah kluster, maka persamaan yang digunakan untuk menghitung sentroid baru adalah:

$$C_i = \frac{1}{M} \sum_{j=1}^M x_j \quad (1)$$

Sedangkan untuk menghitung jarak menggunakan rumus *Euclidean* sebagai berikut:

$$D(x_2, x_1) = \|x_2 - x_1\| = \sqrt{\sum_{j=1}^p |x_{2j} - x_{1j}|^2} \quad (2)$$

Algoritma K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain (Prasetyo, 2012):

Algoritma K-NN adalah sebagai berikut :

1. Menentukan jumlah tetangga terdekat (K)
2. Membagi data dalam data pelatihan dan data uji (data baru)
3. Hitung jarak antara data baru dengan semua data training. Penentuan jarak dapat dilakukan dengan menggunakan rumus *Euclidean*
4. Urutkan jarak dari yang terkecil sampai yang terbesar. Tentukan K buah data yang merupakan tetangga terdekat (mempunyai jarak terdekat)
5. Klasifikasikan data baru tersebut ke dalam kelas/label yang mayoritas

METODE PENELITIAN

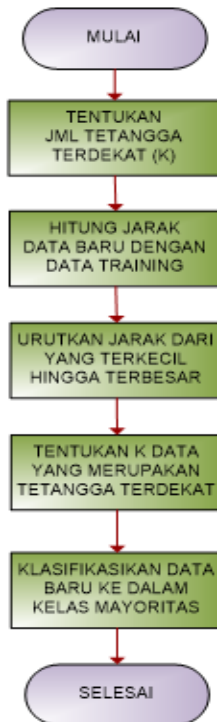
Tahapan penelitian yang dilakukan adalah sebagai berikut:

1) Pengumpulan dan Pengolahan Data

Data yang diperlukan dalam penelitian ini adalah data mengenai parameter yang menentukan kualitas air PDAM meliputi parameter fisik, parameter kimia, dan parameter biologi. Data tersebut diperoleh dari Perusahaan Daerah Air Minum (PDAM) Bandarmasih, Banjarmasin.

2) Perancangan Metode K-Nearest Neighbor (KNN)

Diagram alir metode KNN ditunjukkan dalam Gambar 1.



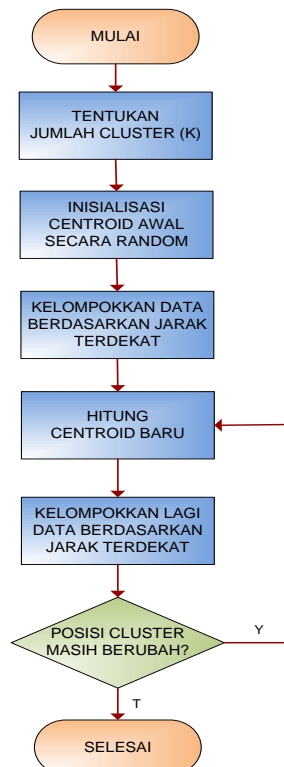
Gambar 1 Diagram Alir Metode KNN

3) Perancangan Metode K-Means

Metode K-Means digambarkan dalam diagram alir seperti ditunjukkan pada Gambar 2.

4) Pembuatan Program Aplikasi dan Desain *User Interface*

Pembuatan program aplikasi ditujukan untuk pengujian sistem sehingga mampu memberikan keputusan terhadap penentuan kualitas air PDAM. Menu program dirancang untuk memudahkan pengguna agar dapat menambahkan data lanjutan di waktu yang akan datang.



Gambar 2 Diagram Alir Metode K-Means

HASIL DAN PEMBAHASAN

Penentuan kualitas air PDAM baik dengan metode KNN maupun K-Means melalui program aplikasi sebagaimana dalam Gambar 3.

Gambar 3 Input Data Uji Pada Proses Klasifikasi

Data uji dimasukkan satu persatu dengan parameternya masing-masing meliputi mikrobiologi, kimia anorganik, parameter fisik, parameter kimia, serta parameter lainnya yang mempengaruhi kualitas air. Kemudian dilakukan pemilihan algoritma yang hendak digunakan yaitu KNN atau K-Means. Proses klasifikasi menggunakan data uji ke-1 dengan metode KNN dan $K = 7$ ditunjukkan dalam Gambar 4, dimana kualitas air yang diuji termasuk dalam kelas Layak.

Hasil Pemeriksaan Data Lengkap		Tambahan																																	
Tanggal Pemeriksaan : 26-08-2019 Minggu Ke : Pertama Bulan : Agustus		Boron	: 0.05 mg/l																																
Mikrobiologi E. Coli : 0 Jlh/100 ml. spi Bakteri Koliform : 0 Jlh/100 ml. spi		Chlorine	: 0.7 mg/l																																
Kimia An-Organik Arsen : 0 mg/l Nitrit : 0.046 mg/l Fluorida : 0.17 mg/l Nitrat : 2.3 mg/l Khromium : 0.01 mg/l Sianida : 0.0035 mg/l		Zat Organik	: 8.85 mg/l																																
Fisik Bau : Tidak Berbau Rasa : Tidak Berasa Warna (TCU) : 15 Suhu : 30 Celsius Kekeruhan (NTU) : 2.06		CO2 Bebas	: 4.4 mg/l																																
Kimia Aluminium : 0.02 mg/l Sulfat : 0 mg/l Besi : 0.128 mg/l Tembaga : 0.02 mg/l Mangan : 0.028 mg/l Ammonium : 0.07 mg/l pH : 7.30 Kesadahan : 72.07 mg/l Seng : 0.2 mg/l Khlorida : 10.64 mg/l		Tabel Ranking <table border="1"> <thead> <tr> <th>Ranking</th> <th>Id data</th> <th>Hasil Analisa</th> <th>Jarak</th> </tr> </thead> <tbody> <tr><td>1</td><td>6</td><td>Layak</td><td>10.3609</td></tr> <tr><td>2</td><td>60</td><td>Layak</td><td>11.6969</td></tr> <tr><td>3</td><td>56</td><td>Layak</td><td>13.3838</td></tr> <tr><td>4</td><td>50</td><td>Layak</td><td>13.7152</td></tr> <tr><td>5</td><td>54</td><td>Layak</td><td>13.9382</td></tr> <tr><td>6</td><td>8</td><td>Layak</td><td>14.5535</td></tr> <tr><td>7</td><td>58</td><td>Layak</td><td>15.6408</td></tr> </tbody> </table>		Ranking	Id data	Hasil Analisa	Jarak	1	6	Layak	10.3609	2	60	Layak	11.6969	3	56	Layak	13.3838	4	50	Layak	13.7152	5	54	Layak	13.9382	6	8	Layak	14.5535	7	58	Layak	15.6408
Ranking	Id data	Hasil Analisa	Jarak																																
1	6	Layak	10.3609																																
2	60	Layak	11.6969																																
3	56	Layak	13.3838																																
4	50	Layak	13.7152																																
5	54	Layak	13.9382																																
6	8	Layak	14.5535																																
7	58	Layak	15.6408																																
Hasil Analisa Nilai K : 7 Total Tidak Layak : 0 Total Layak : 7 Sehingga dapat disimpulkan bahwa air sudah layak konsumsi		Cetak																																	

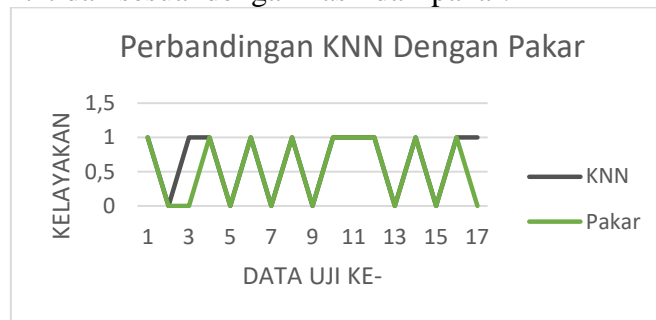
Gambar 4 Perhitungan Data Uji ke-1 Menggunakan KNN

Pengujian dilakukan terhadap seluruh data uji dan kemudian dibandingkan terhadap pakar seperti ditunjukkan dalam Tabel 1. Dari tabel tersebut diperoleh akurasi sebesar 88%.

Tabel 1 Perbandingan KNN Dengan Pakar

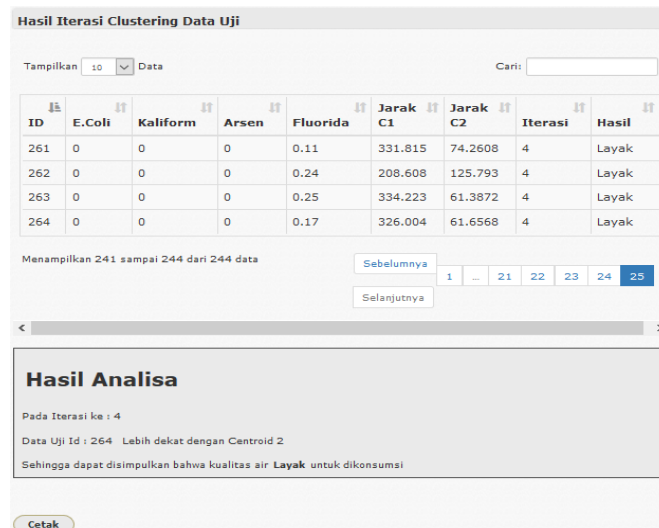
No.	Klasifikasi KNN	Pakar	Hasil Perbandingan
1	Layak	Layak	Sesuai
2	Tidak Layak	Tidak Layak	Sesuai
3	Layak	Tidak Layak	Tidak Sesuai
4	Layak	Layak	Sesuai
5	Tidak Layak	Tidak Layak	Sesuai
6	Layak	Layak	Sesuai
7	Tidak Layak	Tidak Layak	Sesuai
8	Layak	Layak	Sesuai
9	Tidak Layak	Tidak Layak	Sesuai
10	Layak	Layak	Sesuai
11	Layak	Layak	Sesuai
12	Layak	Layak	Sesuai
13	Tidak Layak	Tidak Layak	Sesuai
14	Layak	Layak	Sesuai
15	Tidak Layak	Tidak Layak	Sesuai
16	Layak	Layak	Sesuai
17	Layak	Tidak Layak	Tidak Sesuai

Grafik perbandingan hasil pengujian ditunjukkan dalam Gambar 5 dimana hasil klasifikasi pada data ke-3 dan ke-17 tidak sesuai dengan hasil dari pakar.



Gambar 5 Grafik Perbandingan KNN dan Pakar

Proses perhitungan data uji ke-1 menggunakan algoritma K-Means ditunjukkan pada Gambar 6. Iterasi terus dilakukan sehingga tidak ada lagi data yang berpindah klaster. Gambar 6 menunjukkan pada iterasi ke-4 perhitungan berhenti dan diperoleh data tersebut lebih dekat pada sentroid kedua sehingga kualitas air termasuk dalam kelas Layak.



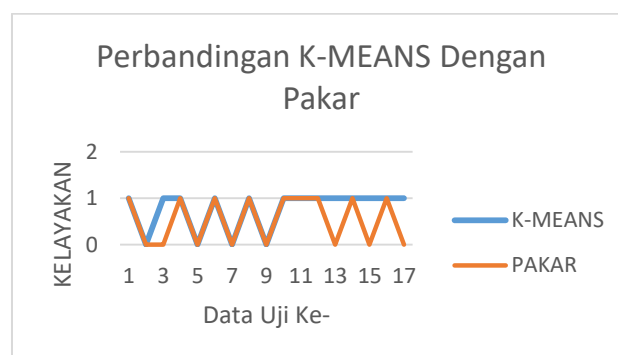
Gambar 7 Hasil Perhitungan K-Means Sampai Iterasi ke-4

Pengujian proses klasifikasi terhadap seluruh data diperoleh akurasi sebesar 76% seperti diberikan dalam Tabel 2.

Tabel 2 Perbandingan K-Means Dengan Pakar

No.	Klasifikasi K-Means	Pakar	Hasil Perbandingan
1	Layak	Layak	Sesuai
2	Tidak Layak	Tidak Layak	Sesuai
3	Layak	Tidak Layak	Tidak Sesuai
4	Layak	Layak	Sesuai
5	Tidak Layak	Tidak Layak	Sesuai
6	Layak	Layak	Sesuai
7	Tidak Layak	Tidak Layak	Sesuai
8	Layak	Layak	Sesuai
9	Tidak Layak	Tidak Layak	Sesuai
10	Layak	Layak	Sesuai
11	Layak	Layak	Sesuai
12	Layak	Layak	Sesuai
13	Layak	Tidak Layak	Tidak Sesuai
14	Layak	Layak	Sesuai
15	Layak	Tidak Layak	Tidak Sesuai
16	Layak	Layak	Sesuai
17	Layak	Tidak Layak	Tidak Sesuai

Grafik perbandingan hasil pengujian ditunjukkan dalam Gambar 8 dimana hasil klasifikasi pada data ke-3, 13, 15, dan ke-17 tidak sesuai dengan hasil dari pakar. Akurasi yang diperoleh menggunakan K-Means lebih rendah dari pada KNN karena K-Means termasuk dalam *unsupervised classification* dimana data yang akan dikelompokkan tidak diketahui kelasnya terlebih dahulu, sedangkan KNN termasuk dalam *supervised classification*.



Gambar 8 Grafik Perbandingan K-Means dan Pakar

KESIMPULAN

Penentuan kualitas air pada penelitian ini menggunakan data dari PDAM Bandarmasih dimana kualitas air yang diuji diklasifikasikan menjadi layak dan tidak layak. Proses klasifikasi menggunakan 27 parameter (atribut) dengan metode KNN dan K-Means. Berdasarkan hasil pengujian diperoleh akurasi dengan KNN lebih baik yaitu 88% dibandingkan dengan K-Means yaitu 76%.

DAFTAR PUSTAKA

- Jabbar, M.A., Deekshatulu B.L, dan Chandra, P., 2013, *Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm*, Procedia Technology 10, 85-94
- Masruro, A., Wibowo, F.W., 2016, *Intelligent Decision Support System For Tourism Planning Using Integration Model Of K-Means Clustering And Topsis*, International Journal of Advanced Computational Engineering and Networking, Volume 4, Issue 1
- Musdalifah, S., Sudarsana, I.W., 2014, *Implementasi Algoritma K-Means Untuk Diagnosa Penyakit Gagal Ginjal Kronis*, Jurnal Ilmiah Matematika dan Terapan, Vol. 11 No. 1
- Nuraziq, A., etc, 2017, *Rancang Bangun Sistem Pendukung Keputusan Penerima Beasiswa dengan Menggunakan Metode Clustering K-Means*, Seminar Nasional Multi Disiplin Ilmu, Vol. 1
- Rahman, M.A., etc, 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
- Sarker, A., etc, 2018, *Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm*, Global Journal of Computer Science and Technology Software & Data Engineering, Vol. 18, Issue 1
- Prasetyo, E., 2012, *Data Mining : Konsep dan Aplikasi Menggunakan MATLAB*, Andi, Yogyakarta