BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Beberapa penelitian yang telah ada sebelumnya yang berkaitan dengan sistem monitoring kualitas air secara *realtime* terangkum dalam tabel 2.1.

Tabel 2. 1 Tabel Referensi

No.	Nama, Tahun	Permasalahan	Data	Kontribusi
1	(Akanksha, 2019)	Kebutuhan air semakin meningkat sementara ketersedian air layak konsumsi menipis, sehingga diperlukan sistem untuk mengukur ketersediaan air pada sebuah penampungan.	Data level/ketinggian air pada penampungan air	Menyajikan strategi manajemen pengendalian air pintar menggunakan teknologi IoT.
2	(Akbar et al., 2019)	Perlunya menjaga kualitas air untuk mengurangi risiko pencemaran yang berdampak pada lingkungan hidup secara *real-time*	Data parameter kualitas air berupa data pH, data suhu, dan data kekeruhan	Menyediakan suatu sistem monitoring kualitas air secara *real-time* berbasis IoT
3	(Handayani. Y, 2020)	pembuangan air limbah domestiknya berakhir di got/selokan/sungai yang mengakibatkan air sungai tercemar	Parameter kualitas air sungai.	Dapat mengklasterkan sungai di surakarta berdasarkan kuaitas airnya
4	(Nurmahaludin, 2019)	Membandingkan hasil uji kualitas air antara model KNN dan K Means pada air PDAM	Data parameter kualitias air PDAM	Mendapatkan metode yang efektif untuk menentukan kualitas air PDAM

No.	Nama, Tahun	Permasalahan	Data	Kontribusi
5	(Savitri & Nursalim, 2023)	Menentukan kualitas air Minum menggunakan Penerapan Algoritma Machine Learning	Data parameter air minum	algoritma Random Forest Classifier memiliki akurasi yang paling baik dibandingkan beberapa algoritma lainnya yang digunakan
6	Susanti, M.D (2025)	Melakukan analisis kualitas air dengan menggunakan Metode K Means Klastering	Data parameter kualitas air berupa pH (keasaman) Turbidity (kekeruhan) Temperature (suhu)	Dapat mengelompokkan kualitas air untuk konsumsi dengan menggunaka metode K Means Klastering

2.2 Dasar Teori

2.2.1 Teori kualitas Air

Air bersih, secara umum diartikan sebagai air yang layak untuk dijadikan air baku bagi air minum. Kelayakan ini bermakna juga bahwa layak untuk mandi, cuci dan kakus. Sebagai air yang layak untuk diminum, bukan berarti langsung dapat diminum, namun harus dimasak hingga mendidih. Menurut Peraturan Menteri Kesehatan RI No. 492/MENKES/PER/IV/2010, air minum adalah air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum. Jenis air minum menurut Peraturan Menteri Kesehatan RI No. 907/MENKES/SK/VII/2002, meliputi: (i) air yang didistribusikan melalui pipa untuk keperluan rumah tangga, (ii) air yang didistribusikan melalui tangki air, (iii) air kemasan, dan (iv) air yang digunakan untuk produksi bahan makanan dan minuman yang disajikan kepada masyarakat. Menurut Hamed (2019), penerapan *K-Means Clustering* pada data kualitas air bertujuan untuk memisahkan kelompok berdasarkan kesamaan parameter dominan seperti pH, kekeruhan (NTU), dan suhu. Hal ini sejalan dengan hasil penelitian, di mana parameter tersebut menjadi penentu utama perbedaan antar klaster.Keruh & Pemanasan dicirikan dengan pH netral (6,96) tetapi suhu cenderung meningkat dan

tingkat kekeruhan tinggi 3,84 NTU. Kondisi ini sesuai dengan teori *K-Means clustering* yang menempatkan sampel dengan ciri serupa dalam satu kelompok (Kung et al.,1992). Secara standar WHO, kekeruhan di atas 1 NTU sudah tidak memenuhi syarat air minum, sehingga klaster ini merepresentasikan kualitas air kurang baik. Sedikit Asam & Stabil memiliki pH 6,69 sedikit di bawah batas aman 6,5–8,5 suhu relatif stabil, dan kekeruhan sedang 2,88 NTU. Menurut penelitian (Darji&Lodha ,2019) klaster seperti ini menggambarkan kondisi air dengan kualitas cukup baik namun memiliki keterbatasan, karena pH mendekati ambang bawah standar.Netral Basa & Stabil ditandai pH 7,23 netral ke basa, suhu stabil, dan kekeruhan sedang (2,89 NTU. Berdasarkan teori PCA + Cluster Analysis (Hamed, 2019), kondisi ini dapat dianggap sebagai klaster terbaik karena pH sesuai standar ideal (6,5–8,5) dan kekeruhan masih relatif rendah. Keruh & Pendinginan menunjukkan pH netral (6,98), suhu cenderung menurun, namun tingkat kekeruhan tinggi (3,87 NTU). Hal ini konsisten dengan studi Jalpa Darji (2025) yang menunjukkan bahwa klaster dengan kekeruhan tinggi selalu masuk kategori kualitas rendah meskipun pH berada dalam rentang normal.

Adapun kualitas air minum yang aman untuk dikonsumsi berdasarkan peraturan Menteri, air harus memenuhi dua parameter standar, yakni wajib dan tambahan. Untuk parameter wajib sendiri, terdiri dari Mikrobalogi, yang artinya, tidak mengandung E-Coli, dan Bakteri Kolfrom. Kemudian, Bebas zat kimia beracun, PH 6,5 sampai 8,5, memiliki TDS Maksimum 500 mg/l, dan suhu maksimal 3 C dari suhu udara

penelitian ini hanya akan menggunakan 3 parameter untuk menentukan kualitas air minum seperti tabel 2.2.

Tabel 2. 2 Tabel Parameter Kualitas Air Minum

Parameter	Satuan	Batas Minimal	Batas Maksimal	Keterangan
Suhu	°C	3	3	±3 dari suhu udara
Kekeruhan	NTU	0	5	
pH	-	6.5	7.5	

Pada tabel 2.2 adalah tabel yang berisi 3 parameter yaitu suhu,kekeruhan,dan ph. Parameter tersebut yang akan digunakan pada analisis kualitas air untuk konsumsi dengan menggunakan metode *K-Means Clustering*.

2.2.2 Algoritma *K Means Clustering*

Data mining adalah suatu proses pengambilan data dari suatu sumber dimana data yang diambil sangatlah besar. Data mining memiliki fungsi descriptive dan predictive. Definisi sederhana dari data mining adalah ekstraksi informasi atau pola yang penting atau enarik dari data yang ada di database yang besar. Dalam jurnal ilmiah, data mining juga dikenal dengan nama *Knowledge Discovery in Databases* (KDD) (Larose & Larose, 2015).

Data mining adalah bagian dari proses KDD yang terdiri dari beberapa tahapan seperti pemilihan data, pra pengolahan, transformasi, data mining, dan evaluasi hasil (Maimon & Last, 2001). Metode data mining adalah cara bagaimana metode tersebut diterapkan. Penerapan metode perlu disesuaikan dengan tujuan penggunanya.

Metode *K-Means Clustering* merupakan salah satu metode klastering dengan partitional, karena *K-Means Clustering* didasarkan pada penentuan jumlah awal kelompok dengan mendefinisikan nilai centroid awalnya (Madhulatha, 2012). Algoritma K-Means berbasis pembagian adalah salah satu jenis algoritma klaster, dan memiliki keunggulan dalam kesederhanaan, efisiensi, dan kecepatan (Li & Wu, 2012).

K-Means Clustering termasuk dalam metode Data Mining *partitioning clustering* yaitu setiap data harus masuk dalam cluster tertentu dan memungkinkan bagi setiap data yang termasuk dalam cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster lain. *K-Means Clustering* memisahkan data ke K daerah bagian terkenal karena kemudian dan kemampuannya untuk mengklasifikasi data besar dan outlier dengan sangat cepat (Siska, 2016).

Tujuan dari *K-Means Clustering* adalah untuk mendapatkan kelompok data dengan memaksimalkan kesamaan karakteristik dalam klaster dan memaksimalkan perbedaan antar klaster.

Algoritma *K-Means Clustering* mengelompokkan data berdasarkan jarak antara data terhadap titik centroid klaster yang didapatkan melalui proses berulang. Analisis perlu menentukan jumlah K sebagai input algoritma. Dalam ranah machine learning, *K-Means Clustering* termasuk ke dalam jenis *unsupervised learning*. Algoritma dari *K-Means Clustering* adalah :

1. Tentukan jumlah kelompok
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat cluster (centroid/rata-rata) dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
5. Kembali ke Langkah 3, apabila masih ada data yang berpindah cluster, atau apabila perubahan nilai centroid ada yang di atas nilai threshold yang ditentukan, atau apabila perubahan nilai pada fungsi obyektif yang digunakan masih di atas nilai threshold yang ditentukan.

Untuk menentukan jarak antara setiap data dengan titik pusat cluster, maka digunakan persamaan Euclidean (Bezdek, 1981) seperti ditunjukkan pada persamaan 1 atau Manhattan/City Block (Miyamoto & Agusta, 1998) yang ditunjukkan pada persamaan 2.

$$D(X_2, X_1) = ||X_2 - X_1||_2 = \sqrt{\sum_{j=i}^{p} |X_{2j} - X_{1j}|^2} \qquad \text{Persamaan (1)}$$

$$D(X_2, X_1) = ||X_2 - X_1||_1 = \sum_{j=i}^{p} X_{2j} - X_{1j}| \qquad \text{Persamaan (2)}$$

Fungsi objektif yang digunakan untuk metode K-Means ditentukan berdasarkan jarak dan nilai keanggotaan data dalam kelompok. Fungsi objektif

dapat ditentukan menggunakan persamaan 3 (MacQueen, 1967). Sedangkan untuk fungsi objektif awal digunakan nilai Absolute.

$$j = \sum_{i=1}^{n} \sum_{i=1}^{k} a_{ic} D(x_1, C)^2$$ Persamaan (3)

2.2.3 Python

Python saat ini menjadi salah satu bahasa pemograman yang banyak digunakan di tahun 2020. Python banyak diminati karena pendekatanya yang ringkas, sederhana dan modular. Python telah digunakan oleh banyak programer dimanis yang mendukung paradigma pemograman berbasis objeck.Keunggunlan python adalah mudah dipahami,mudah dibaca,dan perintah perintah yang ada lebih ringkas dibandingkan bahasa pemograman lainya (Kurniawan & Romzi, 2022). Adapun library yang digunakan dalam bahasa pemograman python yaitu

1. Pandas

   Pandas adalah module pada python yang secara spesifik bekerja untuk mengolah data. Pandas bertujuan untuk menjadi blok penyusun tingakat tinggi yang fundamental untuk, melakukan analisis data dunia nyata yang praktis dalam python.Selain itu pnadas juga memiliki tujuan yang lebih luas untuk menjadi alat analisi/manipulasi data sumber terbuka yang paling kuat dan fleksibel yang tersedia dalam bahasa apapun (Isa Albanna & R. Tri hadi laksono, 2022)
2. Matplotlip

   Matplotlip diciptakan oleh ahli safar john hunter untuk mengolah data. Matplotlib berkembang dan digunakan serta dikembangkan oleh banyak orang di berbagai bidang. Matplotlib adalah pustaka lengkap untuk membuat visualisasi statis,animasi,dan interaktif dalam python. Matplotlib membuat hal hal yang mudah menjadi mudah dan hal hal yang sulit menjadi mungkin. (Kelly Hermanto, 2022)

2.2.4 Google Colabolatory

Penggunaan Google Colaboratory sebagai alat pengantar dasar untuk pemrograman Python. Metodologi penelitian melibatkan analisis fitur-fitur Google Colaboratory, penerapan kode Python dasar, dan pengamatan pengalaman pengguna. Temuan penelitian menunjukkan bahwa Google Colaboratory menyediakan aksesibilitas tinggi tanpa memerlukan persyaratan instalasi atau konfigurasi tambahan, yang memungkinkan pengguna untuk dengan cepat memulai pemrograman Python. Lebih jauh, platform ini menawarkan kemampuan berbagi dan kolaborasi daring serta menyediakan sumber daya komputasi yang cocok untuk tugas-tugas pemrosesan data intensif. Wilyani, F., Arif, Q. N., & Aslimar, F. (2024)

2.2.5 Clustering

Clustering atau klasterisasi adalah suatu teknik atau metode untuk mengelompokkan data.Menurut Tan, 2006 clustering adalah sebuah proses untuk mengelompokkan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum.

Clustering merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan cluster. Hasil clustering yang baik akan menghasilkan tingkat kesamaan yang tinggi dalam satu kelas dan tingkat kesamaan yang rendah antar kelas.Kesamaan yang dimaksud merupakan pengukuran secara numerik terhadapa dua buah objek.Nilai kesamaan antar kedua objek akan semakin tinggi jika deua objek yang dibandingakan memiliki kemiripan yang tinggi ,begitu juga dengan sebaliknya (Salulolo et al., 2016).

2.2.6 Davies-Bouldin Index

Davies bouldin index (DBI) adalah metric untuk mengevaluasi atau mempertimbangkan hasil algoritma clustering. Pertama kali diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Dengan menggunakan DBI suatu cluster akan dianggap memiliki skema clustering yang optimal adalah yang memiliki DBI minimal (Butsianto & Saepudin, 2020). DBI membantu menentukan jumlah klaster optimal dengan menilai kompaknya klaster dan

seberapa jauh klaster satu dengan lainnya. Semakin kecil nilai DBI, semakin baik kualitas klasterisasi.

1. Kompak (Compactness) – S_i
    a. Mengukur seberapa rapat titik titik dalam satu klaster terhadap centroidnya sendiri
    b. Semakin kecil nilai S_i semakin baik, karena klaster lebih padat
2. Pemisahan (Separation) - D_{ij}
    a. Mengukur jarak antara centroid klaster i dan j
    b. Semakin besar D_{ij} semakin baik karena klaster lebih terpisah

DBI untuk setiap klaster i dihitung dengan rumus

$$R_{ij} = \frac{S_i + S_j}{D_{ij}} \qquad \text{Persamaan (4)}$$

Dimana :

$S_i = rata\ rata\ jarak\ titik\ dalam\ klaster\ i\ ke\ centroidnya$

$S_j = rata\ rata\ jarak\ titik\ dalam\ klaster\ j\ ke\ centroidnya$

$D_{ij} = jarak\ antara\ centroid\ klaster\ i\ dan\ j$

Lalu, DBI dihitung sebagai rata rata dari nilai maksimum R_{ij} untuk semua klaster

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{j \neq i}\, R_{ij} \qquad \text{Persamaan (5)}$$

2.2.7 Silhouette score

Silhouette score adalah metrik evaluasi yang digunakan untuk mengukur seberapa baik pemisahan cluster dalam data. Metrik ini memberikan nilai antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa objek dalam sebuah cluster lebih dekat dengan objek dalam cluster yang sama daripada dengan objek dalam cluster lainnya(Saputra & Yusuf, 2024)

Dalam analisis data, khususnya dalam clustering, menentukan jumlah klaster yang optimal merupakan langkah penting untuk mendapatkan hasil yang

akurat dan bermakna. Salah satu metode yang sering digunakan untuk menilai kualitas klasterisasi adalah Silhouette Score. Silhouette Score membantu mengukur seberapa baik suatu data dikelompokkan dalam klaster yang sama, sekaligus seberapa jauh klaster tersebut dari klaster lainnya. Dengan menggunakan metrik ini, kita dapat mengetahui apakah klaster yang terbentuk cukup kompak dan terpisah dengan jelas dari klaster lain.

Silhouette Score adalah metode untuk mengukur kualitas klasterisasi dalam algoritma seperti K-Means. Skor ini menunjukkan seberapa baik titik-titik dalam klaster dikelompokkan dan seberapa jauh klaster satu dengan yang lain. Silhouette Score atau Silhouette Coefficient, memiliki keuntungan yaitu nilai yang dihasilkan dapat digunakan untuk menentukan jumlah cluster alami dalam kumpulan data. Metrik ini merupakan kombinasi dari metode pemisahan dan kohesi (Haq et al., 2023).Untuk mencari Silhouette Score dapat digunakan persamaan
kohesi (Haq et al., 2023). Untuk mencari Silhouette Score dapat digunakan

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad \text{Persamaan (6)}$$

Dimana

$jika\ S(i)\ mendekati\ 1\ klaterisasi\ sangat\ baik$
$jika\ S(i) mendekati\ 0\ titik\ berada\ di\ batas\ antara\ 2\ klaster$
$jika\ S(i)\ negatif\ maka\ titik\ lebih\ dekat\ ke\ klater\ lain\ (klaterisasi\ buruk)$