Deteksi Anomali Pada Segmentasi Konsumsi Air Pelanggan PDAM dengan Metode *K-Means*

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi Teknologi Informasi Kampus Kota Surabaya

Fakultas Informatika

Universitas Telkom

1202200052

Fahmi Ammry Helmi Irfansyah

Program Studi Sarjana Teknologi Informasi (Kampus Kota Surabaya)

Fakultas Informatika

Universitas Telkom

Surabaya

2024

LEMBAR PENGESAHAN

DETEKSI ANOMALI PADA SEGMENTASI KONSUMSI AIR PELANGGAN PDAM DENGAN METODE *K-MEANS*

Anomaly Detection in Water Consumption Segmentation of PDAM Customers With K-Means Method

NIM :1202200052

Fahmi Ammry Helmi Irfansyah

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Teknologi Informasi (Kampus Kota Surabaya)
Fakultas Informatika
Universitas Telkom

Surabaya, 14 Agustus 2024
Menyetujui

Pembimbing I,	Pembimbing II,
Moh. Hamim Zajuli Al Faroby, S.Si., M.Mat.	Mustafa Kamal, S.Kom., M.Kom.
NIP: 23950023	NIP: 22820015

Ketua Program Studi
Sarjana Teknologi Informasi,

Bernadus Anggo Seno Aji, S.Kom., M.Kom.
NIP: 23929009

LEMBAR PERNYATAAN

Dengan ini saya, Fahmi Ammry Helmi Irfansyah, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul **"Deteksi Anomali Pada Segmentasi Konsumsi Air Pelanggan PDAM dengan Metode K-Means"** beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Surabaya, 14 Agustus 2024
Yang Menyatakan

Fahmi Ammry Helmi Irfansyah

Deteksi Anomali Pada Segmentasi Konsumsi Air Pelanggan PDAM dengan Metode *K-Means*

Fahmi Ammry Helmi Irfansyah[1], Moh. Hamim Zajuli Al Faroby [2], Mustafa Kamal [3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Surabaya
[1]fahmiammryhi@students.telkomuniversity.ac.id, [2]alfarobymhz@telkomuniversity.ac.id,
[3]mustafakamal@telkomuniversity.ac.id

Abstrak

Pemakaian air merupakan hal yang penting dalam memenuhi kebutuhan masyarakat, namun seringkali terjadi anomali dalam konsumsi air pelanggan. Anomali terjadi ketika ada penyimpangan dari pola hubungan yang diharapkan antara pemakaian air dan tekanan air PDAM. Dalam rangka mengatasi masalah tersebut dibutuhkan analisis untuk mengetahui pola tren anomali pada data setiap pelanggan. Penelitian ini bertujuan untuk mengidentifikasi konsumsi air pelanggan yang tidak wajar atau anomali, serta melakukan segmentasi pelanggan berdasarkan pola tren konsumsi bulanan menggunakan metode *K-Means*. Dalam penelitian ini, data yang diperoleh berasal dari PDAM dan data yang digunakan adalah data pelanggan yang berada di Surabaya pada tahun 2023, meliputi data lokasi pelanggan, data lokasi sensor, data konsumsi air pelanggan dan data tekanan air. Hasil penelitian menunjukkan bahwa dari total 130.023 data, teridentifikasi 94690 data anomali. Evaluasi menggunakan *silhouette score* menghasilkan nilai 0,9415 menunjukkan bahwa *clustering* berjalan cukup baik. Hasil analisis divisualisasikan melalui website yang dapat memudahkan PDAM dalam menganalisis data dan mengoptimalkan proses bisnisnya.

Kata kunci : Deteksi Anomali, Segmentasi, K-Means, Website, PDAM.

Abstract

Water consumption is important in meeting the needs of society, but there are often anomalies in customer water consumption. Anomalies occur when there is a deviation from the expected pattern of relationship between water usage and PDAM water pressure. In order to overcome this problem, an analysis is needed to determine the anomalous trend pattern in each customer's data. This study aims to identify anomalous customer water consumption, and segment customers based on monthly consumption trend patterns using the K-Means method. In this study, the data obtained comes from PDAM and the data used is customer data located in Surabaya in 2023, including customer location data, sensor location data, customer water consumption data and water pressure data. The results showed that from a total of 130,023 data, 94690 anomalous data were identified. Evaluation using the silhouette score resulted in a value of 0.9415, indicating that the clustering runs quite well. The results of the analysis are visualised through a website that can facilitate the PDAM in analysing data and optimising its business processes.

Keywords: Anomaly Detection, Segmentation, K-Means, Website, PDAM.

1. Pendahuluan

Dalam era saat ini, pemanfaatan teknologi informasi tidak hanya mempermudah pekerjaan manusia, tetapi juga menjadi landasan utama dalam pengelolaan bisnis perusahaan sehingga tujuan atau permasalahan pada pekerjaan tersebut dapat diatasi secara efektif. PDAM Surya Sembada Surabaya merupakan sebuah perusahaan Badan Usaha Milik Daerah (BUMD) di Surabaya yang fokus pada penyediaan air bersih untuk kebutuhan masyarakat [1]. Perusahaan ini juga mengimplementasikan perkembangan teknologi informasi untuk mengelola berbagai kegiatan bisnisnya, termasuk proses pembayaran tagihan air, penanganan keluhan, dan fungsi lainnya. PDAM Surya Sembada Surabaya menghadapi beberapa permasalahan dalam mengidentifikasi dan menangani pola perilaku konsumsi air yang tidak wajar. Permasalahan ini dapat mencakup kendala seperti pembacaan meter yang tidak akurat, variasi konsumsi air yang signifikan di antara pelanggan, dan tantangan lainnya yang dapat mempengaruhi efisiensi operasional dan proses bisnis. Oleh karena itu, melalui penelitian ini, akan dirancang sistem deteksi anomali yang terintegrasi dalam bentuk website.

Dalam konteks penelitian ini, anomali didefinisikan sebagai penyimpangan dari pola hubungan yang diharapkan antara pemakaian air dan tekanan air. Secara normal, hubungan antara pemakaian air dan tekanan air seharusnya berbanding terbalik. Artinya, ketika pemakaian air rendah, tekanan air cenderung tinggi, dan sebaliknya. Hal ini dapat dianalogikan dengan situasi ketika kran dibuka sedikit (pemakaian rendah), tekanan air dari pompa akan naik. Anomali terjadi ketika hubungan ini menyimpang dari pola yang diharapkan, yaitu ketika pemakaian air dan tekanan air justru berbanding lurus. Contohnya, jika terdeteksi pemakaian air yang rendah bersamaan dengan tekanan air yang juga rendah, atau pemakaian air tinggi dengan tekanan air yang juga tinggi, maka situasi tersebut dianggap sebagai anomali.

Dalam Data Mining, deteksi anomali merupakan aspek penting yang membantu menemukan perilaku anomali pada data yang paling rentan menjadi sebuah ancaman [2]. Terkait dengan hal ini, penerapan *K-Means* dalam deteksi anomali pada data pemakaian air pelanggan PDAM Surya Sembada Surabaya dapat meningkatkan keandalan sistem. Segmen yang terbentuk digunakan sebagai acuan untuk mengidentifikasi konsumsi air yang berpotensi menjadi anomali atau di luar pola normal, sehingga tindakan korektif dapat diambil lebih cepat.

Segmen merupakan teknik multivariat dengan tujuan utama untuk mengelompokkan objek berdasarkan karakteristiknya [3]. Dalam penelitian ini, segmentasi pelanggan berdasarkan pola tren konsumsi air menggunakan *K-Means Time Series*, yang merupakan variasi dari metode *cluster*ing *K-Means* untuk data *Time Series*. *K-Means Time Series* dirancang khusus untuk data deret waktu. *K-Means Time Series* adalah pengembangan dari algoritma *K-Means* yang mempertimbangkan aspek temporal data, memungkinkan pengelompokan berdasarkan kesamaan pola konsumsi air dari waktu ke waktu, bukan hanya berdasarkan nilai-nilai statis [4]. Penggunaan algoritma *K-Means Time Series* dalam pengelolaan data pemakaian air pelanggan PDAM dapat membantu mengidentifikasi pola dan perilaku konsumsi air yang tidak wajar atau anomali. Dengan menggunakan K-Means Time Series, PDAM Surya Sembada Surabaya dapat menganalisis data pelanggannya dengan lebih komprehensif, memahami pola konsumsi air yang berubah dari waktu ke waktu, dan mengidentifikasi anomali dengan lebih akurat [5].

Keunggulan dalam penelitian ini adalah penerapan metode *K-Means Time Series* yang mampu mengelompokkan dan memberikan penyelesaian yang tepat dalam mendeteksi anomali pada data deret waktu [6]. Meskipun metode lain seperti DBSCAN untuk data temporal atau *Hierarchical Clustering* juga dapat dipertimbangkan, *K-Means Time Series* dipilih untuk menguji efektivitasnya dalam mendeteksi anomali pada pola konsumsi air pelanggan terhadap tekanan air dari waktu ke waktu. Hasil dari penelitian ini dapat membantu perusahaan PDAM mengidentifikasi pelanggan yang tidak wajar atau anomali pada konsumsi air pelanggan serta menguji efektivitas metode *clustering* dengan *Silhouette Score*.

Penelitian ini memfokuskan pada analisis data dari sistem deteksi anomali untuk membantu perusahaan PDAM dalam proses bisnisnya. Topik yang diangkat adalah bagaimana analisis data dari sistem deteksi anomali dapat digunakan untuk mengidentifikasi pelanggan yang menunjukkan pola konsumsi air yang tidak wajar atau anomali. Sistem deteksi anomali yang digunakan dalam penelitian ini memanfaatkan algoritma *K-Means Time Series*, yang bertujuan untuk mengelompokkan pelanggan berdasarkan pola konsumsi air mereka. Analisis ini diharapkan dapat membantu PDAM dalam mengidentifikasi pelanggan dengan konsumsi air yang anomali sehingga dapat dilakukan tindakan yang tepat dalam pengelolaan distribusi air.

Batasan masalah pada penelitian ini adalah Fitur pada dashboard yang hanya menampilkan fitur Deteksi anomali dan visualisasi *clustering* serta cek status anomali. Pembatasan ini diperlukan mengingat keterbatasan waktu dan sumber daya yang tersedia, serta untuk memastikan bahwa penelitian dapat diselesaikan dalam waktu satu semester. Dengan adanya batasan ini, diharapkan penelitian dapat dilakukan secara efektif dan memberikan hasil yang relevan dan dapat diandalkan.

Luaran ini bertujuan mengidentifikasi anomali konsumsi air pelanggan terhadap tekanan air PDAM menggunakan algoritma *K-Means Time Series*, serta website untuk visualisasi hasilnya. selain itu juga untuk mengetahui performa *clustering* menggunakan *silhouette score* untuk memastikan akurasi dan seberapa efektif metode *K-Means Time Series* dalam mendeteksi anomali.

2. Studi Terkait

Penelitian terkait Dalam setiap proses penelitian, seorang peneliti akan mengidentifikasi dan meninjau penelitian-penelitian sebelumnya yang terkait dengan topik yang diteliti. Proses ini melibatkan perbandingan antara ide baru peneliti dengan temuan yang ada dalam penelitian sebelumnya untuk memperoleh pengaruh dari penelitian sebelumnya. Laporan penelitian yang diusulkan juga akan menyertakan hasil penelitian sebelumnya yang relevan sebagai bagian dari konteks penelitian yang lebih luas.

Penelitian yang dilakukan oleh Amri Muhaimin pada tahun 2018, membahas tentang penggunaan metode *Kohonen SOM* dan *local outlier factor* untuk mendeteksi anomali. Hasil Penelitian menunjukan bahwa Metode *Kohonen SOM* dan *Local Outlier Factor* mampu menangkap pola anomali yang tidak dapat ditangkap dengan menggunakan metode dari PDAM [7].

Penelitian yang dilakukan oleh Satria Ardi Perdana, Sara Famayla Florentin, dan Agus Santoso pada tahun 2022, membahas tentang penggunaan metode *K-Means* untuk segmentasi pelanggan. Hasil Penelitian menunjukan bahwa metode *clustering K-Means* berhasil dengan baik dalam membentuk tiga cluster pelanggan berdasarkan kriteria umur, jenis kelamin, frekuensi pesanan, tipe pembayaran, dan kota pembelian barang selama satu bulan. Segmentasi ini memfasilitasi pemahaman yang lebih akurat terhadap kebutuhan pelanggan setelah divisualisasikan dan diinterpretasikan [11].

Penelitian yang dilakukan oleh Yogi Tri Cahyono pada tahun 2016, membahas tentang Penggunaan metode *Pearson's Correlation* untuk analisis pola abnormal. Hasil Penelitian menunjukan bahwa metode *Pearson's Correlation* dapat digunakan untuk mendeteksi pola penurunan konsumsi air minum pelanggan yang dianggap

abnormal. Namun, analisis data yang dilakukan dalam tugas akhir ini akan lebih baik jika menggunakan rentang data yang lebih singkat [12].

Dari beberapa penelitian sebelumnya, terdapat beberapa peneliti yang menggunakan metode yang berbeda seperti *Kohonen SOM*, *local outlier factor* dan *Pearson's Correlation* yang memiliki keunggulan dalam menangkap pola anomali yang tidak dapat diakses oleh metode tradisional. Meskipun demikian, kelemahan metode tersebut belum sepenuhnya dapat diatasi. Oleh karena itu, *K-Means Time Series*, yang merupakan variasi dari *K-Means* memiliki potensi untuk memberikan kontribusi baru pada deteksi anomali. *K-means Time Series* merupakan metode yang efektif untuk mendeteksi anomali pada data temporal. Proses *clustering* dengan *K-means Time Series* menghasilkan centroid untuk kelompok data normal dan anomali, yang kemudian dapat digunakan untuk mendeteksi anomali dalam data baru dalam konteks yang sama. Centroid ini memungkinkan deteksi anomali dengan perhitungan jarak yang minimal, sehingga metode ini dapat diterapkan secara real-time dan dalam skala besar. Dalam pendekatan ini, *K-means Time Series* digunakan untuk mengidentifikasi cluster yang mewakili data normal dan anomali [16].

2.1 *K-Means Clustering*

K-Means clustering adalah algoritma clustering yang digunakan untuk membagi data menjadi kelompok-kelompok berdasarkan kesamaan karakteristik. Algoritma ini bekerja dengan cara mengelompokkan data ke dalam k klaster, di mana setiap data akan ditempatkan ke *cluster* yang memiliki pusat (centroid) terdekat dengannya. Proses ini dilakukan dengan menghitung jarak antara setiap data dengan pusat *cluster*, dan kemudian memperbarui posisi pusat *cluster* berdasarkan rata-rata dari data yang termasuk dalam *cluster* tersebut. Berikut tahapan menggunakan Metode *K-means*:

1. Tentukan k sebagai kandidat banyak cluster yang akan dibentuk. Elbow Methods digunakan untuk memilih jumlah k-cluster yang akan digunakan untuk pengelompokan data pada algoritma *K-Means* [13].
2. Tentukan titik pusat (centroid) cluster di awal secara acak. Penentuan centroid awal dilakukan secara random dari objek yang tersedia sebanyak k cluster.
3. Menghitung jarak setiap objek ke setiap centroid menggunakan Euclidean distance [8].

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \textbf{(2.1)}$$

d = Jarak Euclidean
x_i = Variabel pada objek x
y_i = Variabel pada objek y
i = indeks Variabel

4. Setiap data dipisahkan berdasarkan kedekatannya dengan centroid melalui pencarian jarak terpendek.
5. Tentukan centroid baru dengan mengambil rata-rata dari setiap klaster yang terkait menggunakan rumus sebagai berikut:

$$ClusterCenter = \sum_{i=1} \frac{a_i}{n} \quad \textbf{(2.2)}$$

$ClusterCenter$ = Pusat Cluster
a_i = Nilai atribut objek dalam cluster
n = Jumlah dalam cluster
i = Indeks objek dalam cluster

6. Jika terdapat perubahan data dalam setiap cluster, maka ulangi langkah 3 hingga 5 hingga tidak ada perubahan dalam anggota pada setiap cluster.

2.2 *Elbow Method*

Metode Elbow adalah teknik yang digunakan untuk menentukan jumlah *k* optimal dalam analisis *clustering*. Metode ini bekerja dengan melakukan perhitungan selisih kuadrat (distortion score) pada beberapa nilai *k* uji yang berbeda. Distortion score merepresentasikan tingkat variasi dalam cluster. Semakin tinggi nilai *k*,

semakin kecil rata-rata tingkat distorsi, menunjukkan anggota cluster yang lebih mirip satu sama lain. Titik di mana penurunan distorsi melambat secara signifikan membentuk siku, mengindikasikan jumlah cluster optimal untuk pengelompokan data yang lebih efektif [9].

2.3 *Silhouette Score*

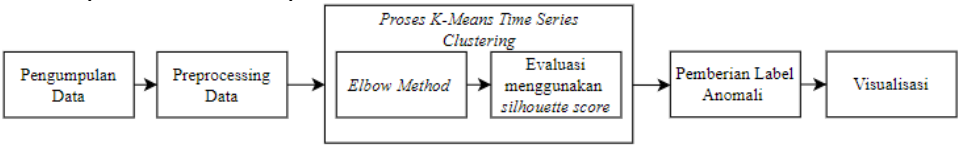
Silhouette score adalah metode evaluasi clustering yang digunakan untuk menilai kualitas *cluster* dengan mengukur seberapa baik objek-objek dikelompokkan dalam *cluster* yang benar. Metode ini mengevaluasi seberapa erat hubungan antara objek dalam satu *cluster* dan seberapa jauh cluster tersebut terpisah dari *cluster* lain. Nilai *silhouette score* berada dalam rentang -1 hingga 1; nilai mendekati 1 menunjukkan bahwa objek-objek dalam *cluster* dikelompokkan dengan baik, sedangkan nilai mendekati -1 menunjukkan bahwa objek-objek dalam *cluster* dikelompokkan dengan buruk [10]. Metode *silhouette score* ini dipilih sebagai metrik evaluasi untuk algoritma *K-Means Time Series* karena karakteristiknya yang sesuai dengan sifat *unsupervised learning* dari *K-Means*. Tidak seperti *Davies-Bouldin Index* (DBI) yang fokus pada perbandingan jarak dalam cluster dan antar cluster, *silhouette score* memberikan gambaran yang lebih menyeluruh dengan mengukur seberapa baik setiap objek berada dalam clusternya sendiri dibandingkan dengan cluster terdekat lainnya. Nilai yang dihasilkan berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan kualitas *clustering* yang lebih baik. Selain itu, *silhouette score* tidak tergantung pada bentuk *cluster*, sehingga lebih fleksibel untuk berbagai jenis dataset yang digunakan dalam *K-Means.* Karakteristik-karakteristik ini menjadikan *silhouette score* pilihan yang lebih tepat untuk mengevaluasi hasil clustering *K-Means Time Series*, memberikan penilaian yang lebih akurat dan informatif tentang kualitas pengelompokan dibandingkan DBI [15].

2.4 Deteksi Anomali

Deteksi anomali adalah suatu metode analisis untuk mengenali keberadaan data atau perilaku yang signifikan atau mencolok di antara data atau perilaku lain dalam suatu konteks tertentu. Dalam dasar teorinya, anomali atau outlier dapat diartikan sebagai titik data yang potensial menyimpang dari pola yang umum atau biasa. Dengan kata lain, outlier dianggap sebagai data yang tidak mengikuti pola umum atau standar dalam suatu himpunan data, sehingga deteksi anomali menjadi kritis untuk mengidentifikasi dan memahami keberadaan data yang tidak sesuai dengan norma yang telah ditetapkan [14].

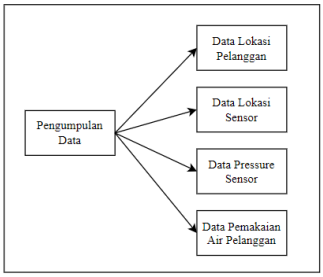
3. Metode

Penelitian ini dilakukan sesuai dengan gambar 1. Langkah pertama adalah pengumpulan data. Setelah itu, dilakukan preprocessing data yang bertujuan untuk normalisasi data dan memfilter data. Proses selanjutnya adalah *K-Means Time Series clustering* dimulai dengan menentukan nilai *k* optimal menggunakan *Elbow Method*. Setelah nilai *k* diperoleh, Evaluasi hasil *clustering* menggunakan *silhouette score* untuk mengetehaui akurasi *clustering*, Terakhir dilakukakan pelabelan anomali pada data dan divisualisasikan melalui website dashboard.



Gambar 1. Alur Penelitian

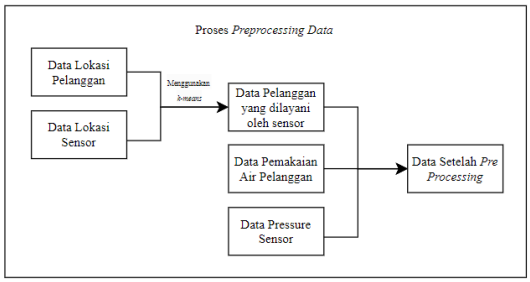
Pada gambar 2. Menggambarkan Pengumpulan Data dilakukan dengan proses permintaan data dari perusahaan PDAM. Terdapat 4 jenis data yang digunakan dalam penelitian ini yaitu data lokasi pelanggan, data pemakaian air pelanggan, data lokasi sensor dan data pressure sensor.



Gambar 2. Pengumpulan Data

Sesuai dengan Gambar 3, alur *preprocessing* dimulai dengan menggabungkan data lokasi pelanggan dan data lokasi sensor. Proses ini menghasilkan data yang menunjukkan pelanggan mana yang dilayani oleh sensor

terdekat. Data pelanggan yang telah diidentifikasi ini kemudian digabungkan dengan data pemakaian air pelanggan serta data pressure sensor yang hasil akhirnya adalah data setelah *preprocessing*.



Gambar 3. Proses *Preprocessing* Data

Pada proses preprocessing, langkah pertama adalah menentukan pelanggan yang dilayani oleh sensor menggunakan metode *Euclidean Distance*. Data yang digunakan meliputi data lokasi pelanggan, dan data lokasi sensor sesuai pada Tabel yang terdapat pada lampiran 1. Dengan pendekatan *Euclidean Distance*, kami dapat mengidentifikasi pelanggan yang dilayani oleh sensor terdekat di sekitar mereka. Pendekatan ini dipilih karena PDAM tidak memiliki data yang spesifik untuk menentukan sensor tekanan mana yang mlayani setiap pelanggan. Pihak PDAM telah memberikan izin untuk menggunakan pendekatan ini. Sebelum menggabungkan data lokasi pelanggan dan data lokasi sensor, dilakukan terlebih dahulu penyaringan data lokasi sensor. Sensor yang memiliki nama IPAM, RP, RESERVOIR, OUTLET, atau INLET akan dihapus dari data karena bukan merupakan pompa dari PDAM.

Pada tahap *preprocessing*, dilakukan penggabungan antara data lokasi pelanggan, dan data lokasi sensor yang terdapat pada lampiran 1. Hasil dari penggabungan ini akan menghasilkan data pelanggan yang dilayani sensor yang terdapat pada lampiran 1, data tersebut sangat penting karena akan menjadi dasar dalam analisis lebih lanjut. Setelah itu dilakukan penggabungan antara data pelanggan yang dilayani sensor, dengan data pemakaian air pelanggan, serta data pressure sensor yang terdapat pada lampiran 1. Pada tabel data pemakaian air pelanggan yang terdapat pada lampiran 1 berisikan Nomor pelanggan, Tahun, Bulan dan Pemakaian yang merupakan konsumsi air pelanggan dengan satuan meter kubik. Pada tabel data pressure sensor pada lampiran 1 Berisikan Id sensor sebagai perangkat sensor tekanan air, Waktu, dan pressure yang merupakan tekanan air dengan satuan bar.

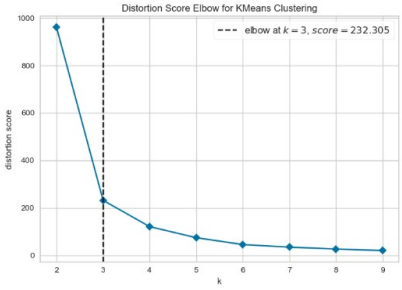
Kombinasi ketiga tabel ini memungkinkan analisis yang menyeluruh dengan mempertimbangkan korelasi antara konsumsi air dan tekanan air, yang sangat penting dalam identifikasi anomali. Dimensi temporal yang ada pada tabel memfasilitasi analisis tren dan pola konsumsi air dari waktu ke waktu. Selain itu, adanya identifikasi unik melalui Nomor Pelanggan dan Id Sensor memungkinkan penghubungan data secara spesifik, meningkatkan akurasi dalam deteksi anomali. Dengan menggunakan ketiga tabel ini, penelitian dapat menghasilkan analisis yang lebih mendalam dan akurat, mencakup baik perspektif pelanggan maupun aspek teknis distribusi air, sehingga memberikan pemahaman yang lebih komprehensif tentang pola konsumsi air dan potensi anomali yang mungkin terjadi.

Pada *preprocessing* Data juga dilakukan dengan membersihkan data yang kosong atau *missing value*, yaitu sebanyak 1 data yang kosong yang terdapat pada data lokasi pelanggan. Selanjutnya, Data Pelanggan yang dilayani Sensor yang berjumlah 629.458 Data digabungkan dengan Data Pelanggan yang berjumlah 7.347.017 data lalu digabungkan lagi dengan Data Sensor yang berjumlah 1.988.475 data, sehingga menghasilkan data yang siap digunakan. Proses ini juga melibatkan penyesuaian waktu data sensor dengan data pelanggan. Selanjutnya, dilakukan normalisasi pada nilai Pressure dari data sensor yang merupakan tekanan air atau pressure dengan satuan bar, dan pemakaian yang dengan satuan meter kubik menggunakan MinMaxScaler. MinMaxScaler dipilih karena metode ini efektif dalam mengubah skala data ke rentang [0, 1], sehingga memudahkan proses clustering dengan algoritma *K-Means*. Normalisasi ini memastikan bahwa semua variabel berada dalam skala yang sama, menghindari dominasi variabel dengan skala yang lebih besar pada hasil *clustering*. Setelah itu, data yang telah di *preprocessing* menghasilkan data yang siap digunakan seperti pada Tabel 1.

Tabel 1. Sampel Data Setelah Preprocessing

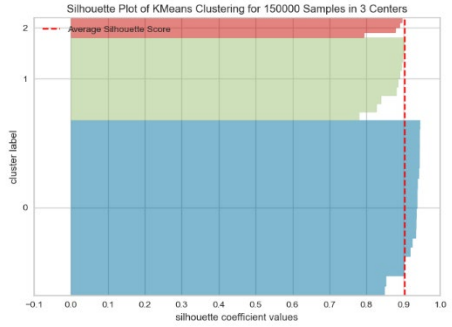
Nomor Pelanggan	Device Id	TAHUN	BULAN	Pressure	Pemakaian
52610000	0f810610-9c64-11ed-82d7-39440ef6fe5d	2023	3	0.0274	0.0001
52610000	0f810610-9c64-11ed-82d7-39440ef6fe5d	2023	4	0.0276	0.0001
52610000	0f810610-9c64-11ed-82d7-39440ef6fe5d	2023	5	0.0280	0.0001
52610000	0f810610-9c64-11ed-82d7-39440ef6fe5d	2023	6	0.0281	0.0001
52610000	0f810610-9c64-11ed-82d7-39440ef6fe5d	2023	7	0.0280	0.0001

Elbow Method digunakan untuk menentukan nilai *k* yang optimal dengan memplot distortion score terhadap berbagai nilai *k*. Distortion score mengukur total jarak kuadrat antara setiap titik data dan centroid *cluster*-nya, di mana penurunan nilai distortion score menunjukkan bahwa data lebih terkelompok dengan baik. Pada grafik *Elbow Method*, titik "elbow" adalah tempat di mana penurunan distortion score mulai melambat secara signifikan, menunjukkan bahwa penambahan *cluster* lebih lanjut tidak memberikan peningkatan substansial dalam kualitas *clustering*. Hal ini mengindikasikan bahwa ada jumlah *cluster* tertentu yang optimal untuk model ini, seperti yang terlihat pada gambar 4, yang menunjukkan nilai *k* yang optimal.



Gambar 4. *Elbow Method*

Setelah menentukan *k* optimal, proses *clustering* dilakukan menggunakan *K-Means Times Series* dengan jumlah *cluster* ini. *K-Means Time Series clustering* menggunakan metrik Euclidean untuk mengukur jarak antara titik data dalam ruang fitur. Metrik ini dipilih karena kesederhanaannya dalam mengukur jarak linier antar titik data, yang cocok untuk data time series yang telah dinormalisasi dan diproses. Hasil *clustering* dievaluasi menggunakan *silhouette score*, yang mengukur seberapa baik setiap titik data cocok dengan *cluster*-nya dibandingkan dengan *cluster* lain. Nilai *silhouette score* yang tinggi menunjukkan bahwa titik data berada pada *cluster* yang benar, sedangkan nilai rendah menunjukkan kekurangan dalam pemisahan antar *cluster*. Evaluasi ini membantu memastikan kualitas *clustering* dan identifikasi pola konsumsi air yang anomali, sesuai pada gambar 5 yang menunjukkan grafik *silhouette score*.



Gambar 5. *Silhouette Score*

Setelah tahap clustering dengan algoritma *K-Means Time Series* selesai dan nilai *k* optimal telah ditentukan, proses berikutnya adalah pemberian label anomali pada data. Pemberian label ini bertujuan untuk mengidentifikasi data yang menunjukkan perilaku tidak wajar. Deteksi anomali dilakukan dengan menganalisis pola tren pelanggan pada setiap *cluster*. *Cluster* yang menunjukkan hubungan langsung atau berbanding lurus antara variabel dianggap sebagai anomali. Sebaliknya, *cluster* yang menunjukkan hubungan terbalik antara variabel dianggap sebagai data normal. Pendekatan ini memungkinkan identifikasi anomali berdasarkan pola hubungan yang tidak sesuai dengan tren umum dalam *cluster*. Oleh karena itu, *cluster* yang menunjukkan hubungan berbanding lurus dilabeli sebagai anomali, sedangkan cluster yang menunjukkan hubungan berbanding terbalik dilabeli sebagai data normal.

Dalam sistem ini, visualisasi hasil deteksi anomali disajikan melalui dua dashboard terpisah yaitu dashboard admin dan dashboard user. Untuk dashboard admin, Proses dimulai dengan meng-upload data CSV yang berisi informasi pelanggan seperti hasil preprocessing. Setelah di-upload, sistem memproses data dan mendeteksi anomali menggunakan algoritma *K-Means Time Series Clustering*. Dashboard admin dirancang untuk mendeteksi anomali pada seluruh data, memberikan visualisasi plot *clustering*, dan menyajikan hasil data CSV yang telah di-*cluster* dan dilabeli anomali yang dapat diunduh. Dashboard ini juga menampilkan total anomali dan total data normal dari file CSV yang dianalisis. Sementara itu, dashboard user dirancang lebih sederhana dan personal, memungkinkan pelanggan individual untuk memeriksa status konsumsi air mereka, melihat apakah terdeteksi sebagai anomali, dan mendapatkan saran atau rekomendasi. Pengguna dapat memasukkan nomor pelanggan dan ID perangkat untuk mendapatkan hasil deteksi anomali. Jika pelanggan terdeteksi sebagai anomali, sistem memberikan saran atau rekomendasi berdasarkan analisis.

4. Hasil dan Pembahasan

Hasil dari deteksi anomali menunjukan bahwa anomali terdapat pada *cluster* 0 dengan total anomali sebanyak 94.690 pelanggan. Preprocessing menghasilkan 130.023 data, Pada proses *Clustering K-Means*, *k* yang di dapat pada penggunaan *Elbow Method* yang ditunjukan gambar 2 mendapatkan nilai optimal *k* = 3. Grafik elbow pada gambar 4, menunjukkan penurunan yang signifikan pada sudut elbow, mengindikasikan bahwa penambahan *cluster* setelah *k* = 3 tidak memberikan peningkatan yang berarti dalam menjelaskan variasi data. Lalu dilakukan evaluasi *silhouette score* setelah proses *clustering* dengan nilai 0,9415. Setelah itu, hasil dari clustering melabeli data berdasarkan analisis pola. Dari 130.032 total data pelanggan, sebanyak 94.690 data pelanggan dilabeli sebagai anomali, sementara 35.333 data pelanggan dilabeli sebagai data normal. Hasil pengclusteran deteksi anomali ini ditampilkan sesuai pada Tabel 2.

Tabel 2. Sampel Hasil Pelabelan Data Setelah *Clustering*

Pressure	Pemakaian	Cluster	Anomali
0.5907	0.0001	2	normal
0.6256	0.0002	2	normal
0.0533	0.0007	0	anomali
0.6931	0.0001	2	normal
0.1136	0.0005	0	anomali

Hasil *Clustering* pada algoritma *K-Means Time Series Clustering* cukup optimal, berdasarkan evaluasi *Silhouette Score* dengan nilai 0,9415 pada algoritma *K-Means Time Series*. *Silhouette Score* sebesar 0,9415 menandakan bahwa *Clustering* berjalan baik. Dengan hasil pengujian yang telah dilakukan sesuai dengan tujuan yang dituju yaitu algoritma *K-Means Time Series Clustering* dapat mengidentifikasi anomali dengan cukup baik.

Hasil analisis menunjukkan adanya tiga cluster yang terbentuk dari data konsumsi air dan tekanan air pelanggan PDAM. Cluster 0, yang dilabeli sebagai anomali, memiliki karakteristik yang berbeda dibandingkan dengan dua cluster lainnya. Pada cluster 0 (anomali), rata-rata pemakaian air yang telah dinormalisasi adalah 0.0002, yang merupakan nilai terendah di antara ketiga cluster. Ini mengindikasikan bahwa pelanggan dalam cluster ini cenderung memiliki tingkat konsumsi air yang relatif rendah. Namun, yang menarik adalah tekanan air rata-rata pada cluster ini juga sangat rendah, yaitu 0.0655. Kombinasi konsumsi air rendah dengan tekanan air yang juga rendah ini tidak sesuai dengan pola normal yang diharapkan, di mana tekanan air seharusnya lebih tinggi ketika konsumsi rendah. Hal ini mengonfirmasi bahwa cluster 0 memang menunjukkan pola anomali.

Sebaliknya, cluster 1 dan 2 yang dilabeli sebagai normal menunjukkan pola yang lebih konsisten dengan ekspektasi. Cluster 1 memiliki rata-rata pemakaian air 0.0003 dan cluster 2 memiliki rata-rata 0.0003, keduanya lebih tinggi dari cluster 0. Yang lebih penting, tekanan air pada kedua cluster ini jauh lebih tinggi, dengan nilai rata-rata sekitar 0.673 untuk keduanya. Hal ini menunjukkan hubungan yang lebih normal antara konsumsi air dan tekanan air, di mana tekanan air tetap tinggi meskipun ada variasi dalam tingkat konsumsi. Perbedaan yang signifikan dalam tekanan air antara cluster anomali 0.0655 dan cluster normal sekitar 0.673 menunjukkan bahwa tekanan air menjadi faktor kunci dalam mengidentifikasi anomali. Pola ini sesuai dengan ekspektasi bahwa anomali dalam konsumsi air seringkali terkait dengan masalah tekanan air yang tidak normal. Hasil dari implementasi sistem deteksi anomali dapat dilihat melalui representasi website dapat dilihat pada lampiran 2.

5. Kesimpulan

Kesimpulan dari penelitian ini menunjukkan bahwa metode *K-Means* berhasil mengidentifikasi konsumsi air pelanggan yang anomali dengan efektif. Dari total 130.032 data, teridentifikasi 94.690 data anomali, memberikan gambaran jelas tentang pola konsumsi air yang tidak wajar. Segmentasi pelanggan menggunakan *K-Means Time Series* menghasilkan 3 *cluster* berbeda berdasarkan karakteristik konsumsi air dan tekanan air. Efektivitas metode ini dibuktikan dengan hasil *Silhouette Score* sebesar 0,9415 menunjukkan kualitas *clustering* yang cukup baik.

Visualisasi hasil melalui dashboard admin dan user memudahkan PDAM dalam menganalisis dan memanfaatkan hasil deteksi anomali, sehingga dapat mengoptimalkan proses bisnisnya melalui identifikasi anomali dan segmentasi pelanggan. Dengan hasil penelitian ini, PDAM dapat mengambil keputusan yang lebih tepat berdasarkan analisis *clustering* yang telah dilakukan. Untuk penelitian selanjutnya, disarankan untuk menggunakan data uji yang lebih besar guna mengoptimalkan nilai cluster, serta melakukan eksplorasi dan perbandingan dengan metode yang lain untuk meningkatkan hasil yang lebih baik.

Daftar Pustaka

[1] Mediana, D., & Nurhidayat, A. I. (2018). Rancang Bangun Aplikasi Helpdesk (A-Desk) Berbasis Web Mengunakan Framework Laravel (Studi Kasus Di PDAM Surya Sembada Kota Surabaya. Jurnal Manajemen Informatika, 8(2).

[2] Gadal, S., Mokhtar, R., Abdelhaq, M., Alsaqour, R., Ali, E. S., & Saeed, R. (2022). Machine Learning-Based Anomaly Detection Using K-Mean Array and Sequential Minimal Optimization. Electronics (Switzerland), 11(14). https://doi.org/10.3390/electronics11142158

[3] Gading Sadewo, M., Eriza, A., Perdana Windarto, A., & Hartama, D. (2018). Seminar Nasional Teknologi Komputer & Sains (SAINTEKS) Algoritma K-Means Dalam Mengelompokkan Desa/Kelurahan Menurut Keberadaan Keluarga Pengguna Listrik dan Sumber Penerangan Jalan Utama Berdasarkan Provinsi (Vol. 01). https://www.bps.go.id.

[4] Zhang, Z., Wang, C., Peng, X., Qin, H., Lv, H., Fu, J., & Wang, H. (2021). Solar Radiation Intensity Probabilistic Forecasting Based on K-Means Time Series Clustering and Gaussian Process Regression. IEEE Access, 9. https://doi.org/10.1109/ACCESS.2021.3077475

[5] Eno Ketherin, B., Anjani Arifiyanti, A., Sodik, A., Sistem Informasi, J., & Teknologi Adhi Tama Surabaya, I. (2018). Analisa Segmentasi Konsumen Menggunakan Algoritma K-Means Clustering.

[6] Mudakir, Ahmad Turmudi Zy, & Aswan S. Sunge. (2023). Penerapan Data Mining Untuk Klasifikasi Pengangkatan Karyawan Menggunakan Algoritma K-Means. Jurnal Informatika Teknologi Dan Sains (Jinteks), 5(3). https://doi.org/10.51401/jinteks.v5i3.3369

[7] Muhaimin, A. (2018). Deteksi Anomali Pada Pemakaian Air Pelanggan PDAM Surya Sembada Kota Surabaya Menggunakan Algoritma Kohonen Self Organizing Maps (SOM) dan Local Outlier Factor (LOF).

[8] Putra, B. Y., Azzahra, F. Y., & Erlanda, I. A. (2023). Klasterisasi Pengunjung Mall Menggunakan Algoritma K-Means Berdasarkan Pendapatan dan Pengeluaran. Jurnal Informatika Dan Teknik Elektro Terapan, 11(3s1). https://doi.org/10.23960/jitet.v11i3s1.3392

[9] Permadi, V. A., Tahalea, S. P., & Agusdin, R. P. (2023). K-Means and Elbow Method for Cluster Analysis of Elementary School Data. Progres Pendidikan, 4(1). https://doi.org/10.29303/prospek.v4i1.328

[10] Setiawan, F. A., Sadikin, M., & Kaburuan, E. R. (2022). Analisis Permasalahan Perangkat Pencetak Menggunakan Metode Algoritma K-Means dan K-Medoids. Teknika, 11(2). https://doi.org/10.34148/teknika.v11i2.471

[11] Perdana, S. A., Florentin, S. F., & Santoso, A. (2022). Analisis Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Aplikasi Alfagift. Sebatik, 26(2). https://doi.org/10.46984/sebatik.v26i2.1991

[12] Cahyono, Y. T. (2016). Analisis Pola Abnormal Konsumsi Air Minum Pelanggan PDAM Surya Sembada Surabaya Menggunakan Metode Pearson's Correlation, Abnormally Low Consumption, dan Windowed Analysis.

[13] Ma'ali, A. A., Girinoto, Ghiffari, M. N., & Hadiprakoso, R. B. (2022). Analisis Log Web Server dengan Pendekatan Algoritme K-Means Clustering dan Feature Importance. Info Kripto, 16(3). https://doi.org/10.56706/ik.v16i3.60

[14] Rijal Kamal, M., & Andri Setiawan, M. (2021). Deteksi Anomali dengan Security Information and Event Management (SIEM) Splunk pada Jaringan UII.

[15] Siregar, H. L., Zarlis, M., & Efendi, S. (2023). Cluster Analysis using K-Means and K-Medoids Methods for Data Clustering of Amil Zakat Institutions Donor. *Jurnal Media Informatika Budidarma*, *7*(2).

[16] Münz, Gerhard, Sa Li, and Georg Carle. "Traffic anomaly detection using k-means clustering." Gi/itg workshop mmbnet. Vol. 7. No. 9. 2007.

Lampiran

Lampiran 1. Lampiran Tabel

Tabel Sampel Data Lokasi Pelanggan

Nomor Pelanggan	POS_LAT	POS_LONG
52610000	-7.258732	112.654296
41110000	-7.230093	112.721187
31410000	-7.212942	112.768899
22810000	-7.236560	112.781788
9460495	-7.265236	112.754250

Tabel Sampel Data Lokasi Sensor

Id Sensor	Nama Sensor	Latitude	Longitude
0a8826e0-9c5d-11ed-82d7-39440ef6fe5d	BARUK UTARA	-7.312239	112.780476
0b626c30-9c5b-11ed-82d7-39440ef6fe5d	KALISUMO BOSEM	-7.302371	112.754806
0b528320-9c5a-11ed-82d7-39440ef6fe5d	OUTLET RP KETEGAN	-7.348088	112.711601
0d5881c0-9c59-11ed-82d7-39440ef6fe5d	JEMURSARI	-7.326729	112.735576
0eb7df20-9c63-11ed-82d7-39440ef6fe5d	KUPANG JAYA	-7.274616	112.700905

Tabel Sampel Data Pemakaian Air Pelanggan

Nomor Pelanggan	Tahun	Bulan	Pemakaian
0001070	2023	1	30
0001094	2023	1	10
0001095	2023	1	10
0001096	2023	1	206
0001103	2023	1	10

Tabel Sampel Data Pressure Sensor

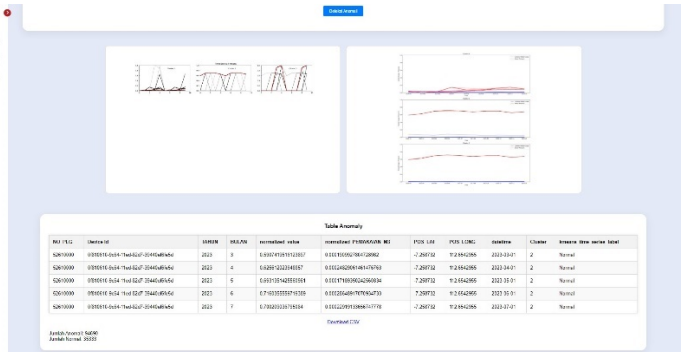
Id Sensor	Waktu	Pressure
01c70a80-6327-11ee-bfca-5d26ac76faaf	2023-12-31 00:00:00	0.4833
01c70a80-6327-11ee-bfca-5d26ac76faaf	2023-12-30 23:00:00	0.4664
01c70a80-6327-11ee-bfca-5d26ac76faaf	2023-12-30 22:00:00	0.4312
01c70a80-6327-11ee-bfca-5d26ac76faaf	2023-12-30 21:00:00	0.4055
01c70a80-6327-11ee-bfca-5d26ac76faaf	2023-12-30 20:00:00	0.3843

Tabel Sampel Data Pelanggan dilayani Sensor

Nomor Pelanggan	Device Id
52610000	0f810610-9c64-11ed-82d7-39440ef6fe5d
41110000	12d02100-ff80-11ed-8f59-45dc1b2f5352
31410000	d5e96670-9c60-11ed-82d7-39440ef6fe5d
22810000	50524af0-9c5f-11ed-82d7-39440ef6fe5d
9460495	ccbe1720-ff80-11ed-8f59-45dc1b2f5352

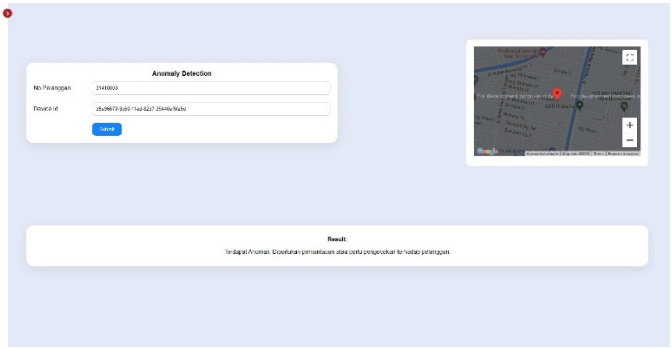
Lampiran 2. Gambar Dashboard

Dashboard Admin dirancang untuk proses deteksi anomali dengan menggunakan visualisasi berbasis website. Hasil deteksi anomali dipresentasikan dalam bentuk plot *cluster* dan tabel hasil deteksi anomali. Pada Gambar Dashboard Admin, tabel ini menunjukkan hasil pelabelan data pelanggan setelah proses clustering, dengan lima entri teratas yang dilabeli sebagai anomali atau normal.



Gambar Dashboard Admin

Pada Dashboard User, pengguna dapat memeriksa status pelanggan mereka untuk mengetahui apakah terdapat anomali atau tidak. Jika terdapat anomali, sistem akan memberikan saran atau rekomendasi berdasarkan analisis. Hal ini dapat dilihat pada Gambar Dashboard User



Gambar Dashboard User