ANALISA DATA KUALITAS AIR KOLAM IKAN AIR TAWAR "PONDSDATA"

Wahyu Hariyadi Nurdin dan Sulistyo Puspitodjati*

Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok, Jawa Barat, 16424
wahyunurdin.whn@gmail.com, sulistyo@staff.gunadarma.ac.id
* Corresponding Author

ABSTRAK

Kualitas air merupakan faktor penting dalam budidaya ikan karena berpengaruh terhadap kesehatan, pertumbuhan, dan produktivitas. Pemanfaatan sensor berbasis Internet of Things (IoT) memungkinkan pemantauan kualitas air secara real-time, namun data yang dihasilkan sering mengandung missing value, outlier, maupun ketidaksesuaian sehingga diperlukan tahapan preprocessing agar analisis lebih akurat. Penelitian ini bertujuan untuk menganalisis kualitas data parameter kualitas air pendekatan Cross-Industry Standard Process for Data Mining (CRISP-DM) pada tahap business understanding, data understanding, dan data preparation. Dataset yang digunakan dari Kaggle dengan judul Pondsdata memliki jumlah baris 74.796 data dengan atribut utama meliputi nitrat, amonia, pH, suhu, dan dissolved oxygen (DO). Tahapan data preparation dilakukan melalui select data, imputasi missing value dengan median dan modus, penanganan outlier menggunakan metode statistik, serta transformasi distribusi dengan normalisasi. Hasil analisis menunjukkan bahwa kadar nitrat memiliki rata-rata 27,16 ppm, amonia 0,0640 mg/L, suhu 26,79°C, DO 10,76 mg/L, dan pH 6,54. Analisis korelasi mengidentifikasi hubungan positif antara suhu dan amonia (0,33), serta hubungan negatif antara pH dan suhu (-0,24) dan antara amonia dan DO (-0,25). Proses clustering menggunakan algoritma K-Means menghasilkan lima klaster dengan karakteristik kualitas air yang berbeda, mencerminkan variasi kondisi perairan dalam dataset untuk ikan nila, ikan mas, ikan patin, ikan gurami, dan ikan lele.

Kata Kunci: *CRISP DM, kualitas air.*

PENDAHULUAN

Kualitas air merupakan faktor kunci dalam budidaya ikan karena memengaruhi kesehatan, pertumbuhan, efisiensi pakan, serta imunitas ikan. Pada sistem *Recirculating Aquaculture System* (RAS), atribut seperti suhu, oksigen terlarut, pH, amonia, dan nitrat menjadi indikator utama kualitas air [1]. Pemantauan kualitas air berbasis sensor memungkinkan pengukuran secara *real-time* dan efisien, namun rentan menghasilkan data tidak valid akibat gangguan teknis, kesalahan kalibrasi, maupun keter-batasan perangkat. Kondisi tersebut dapat menurunkan kualitas dataset dan berimplikasi pada ketidakakuratan analisis [2].

Untuk mengatasi permasalahan ter-sebut, pendekatan *Cross Industry Standard Process for Data Mining* (CRISP-DM) dapat digunakan, khususnya pada tahap *business understanding*, *data understanding*, dan *data preparation* yang berperan dalam meningkatkan kualitas dataset sebelum proses pemodelan *machine learning* [3]. Penelitian mengenai *preprocessing* dan *clustering* telah banyak dilakukan pada berbagai domain, termasuk salah satunya adalah penelitian yang menekankan pentingnya transformasi data numerik menjadi interval dalam tahap *preprocessing* sebelum proses *clustering* menggunakan *K-Means* [4]. Praproses data dengan melakukan normalisasi data dan pembuatan fitur baru dapat meningkatkan kualitas hasil *clustering* [5]. Pendekatan yang sering digunakan dalam mengektrak pengetahuan dari suatu dataset, selain dengan CRISP DM, adalah pendekatan *Knowledge Discovery in Databases* (KDD). Pendekatan KDD, juga melakukan praproses data sebelum proses ekstraksi pengetahuan itu sendiri. Praproses tersebut mencakup seleksi data, pembersihan data, dan transformasi sebelum *clustering* untuk mengelompokkan jenis ikan unggulan, meskipun hasil klaster tidak dikorelasikan dengan standar kualitas lingkungan perikanan [6].

Berdasarkan kajian tersebut, dapat disimpulkan bahwa penelitian terdahulu telah menegaskan bahwa *preprocessing* berperan

penting dalam meningkatkan kualitas *clustering*. Namun, sebagian besar penelitian hanya berfokus pada proses teknis *preprocessing* dan evaluasi klaster tanpa mengaitkan hasil *clustering* dengan standar domain spesifik. Selain itu, penentuan jumlah klaster optimal umumnya hanya mengandalkan satu metrik evaluasi, sehingga justifikasi hasil masih terbatas. Di sisi lain, penelitian kualitas air untuk akuakultur membutuhkan interpretasi berbasis standar biologis, seperti toleransi kualitas air menurut Boyd (1998) [7], agar hasil *clustering* dapat dioptimalkan sebagai dasar pengambilan keputusan budidaya.

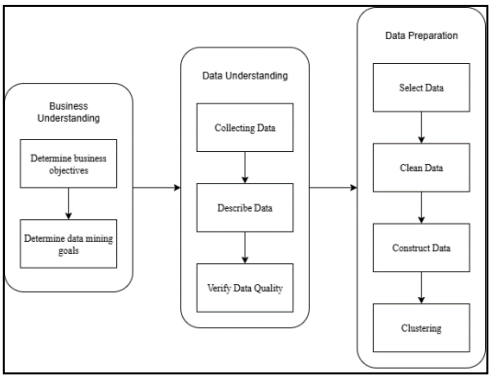
Berbeda dengan penelitian terdahulu, penelitian ini tidak hanya menerapkan tahapan *preprocessing* dan *clustering*, tetapi juga mengevaluasi hasil klaster menggunakan tiga metrik validasi, yaitu WCSS, *Silhouette Score*, dan *Davies–Bouldin Index*. Selain itu, hasil akhir interpretasi klaster dihubungkan dengan standar kualitas air SNI dan toleransi kualitas air berdasarkan Boyd (1998), sehingga memberikan kontribusi yang lebih komprehensif baik dari sisi metodologis maupun aplikatif dalam pengembangan sistem AIoT untuk rekomendasi spesies ikan air tawar.

Penelitian mengenai kualitas air kolam ikan tidak hanya penting bagi bidang akuakultur, tetapi juga relevan dengan pengembangan sistem cerdas berbasis AIoT. Dengan semakin banyaknya data sensor yang tersedia secara daring, tantangan utama bukan lagi pengumpulan data, tetapi bagaimana mengolah dan meningkatkan kualitas data tersebut agar dapat digunakan untuk *machine learning* dan sistem pengambilan keputusan otomatis. Oleh karena itu, penelitian ini menekankan aspek metodologis dalam tahapan *data preprocessing* menggunakan pendekatan CRISP-DM untuk memastikan kesiapan data sebelum digunakan dalam model kecerdasan buatan.

Hasil dari analisa data ini, termasuk hasil *clustering* didalamnya, diharapkan dapat menjadi pertimbangan dalam membangun model prediksi atau ektraksi pengetahuan lainnya, yang merupakan bagian dari pengembangan sistem AIoT Model Pemilihan Spesies Ikan Berdasarkan Atribut Kualitas Air. Dengan demikian, penerapan tahapan *data preprocessing* diharapkan mampu memperkuat fondasi data dan meningkatkan akurasi analisis untuk mendukung keberhasilan sistem yang akan dibangun.

METODE PENELITIAN

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan metode standar dalam proses data mining untuk menyelesaikan permasalahan bisnis yang memiliki fase yang jelas, terstruktur, dan terdokumentasi dengan baik sehingga mudah digunakan [3]. Metode ini fleksibel terhadap adaptasi dan dapat digunakan di berbagai industri untuk mengubah proyek data mining menjadi lebih mudah, cepat, dapat digunakan berulang kali, serta mudah dikelola [8]. CRISP-DM terdiri dari enam fase yang dimulai dari *business understanding* untuk memahami tujuan dari sisi bisnis, kemudian *data understanding* untuk mengenali karakteristik data, *data preparation* untuk menyiapkan data untuk proses *modeling*, fase *evaluation* untuk memastikan kesesuaian model dengan tujuan bisnis, dan terakhir *deployment* untuk mengimplementasikan hasilnya [8]. Dalam implementasinya, waktu CRISP-DM banyak dihabiskan dalam tahap *data preparation* yaitu 50-70%, *data understanding* sebesar 20-30%, *modeling*, evaluasi, dan *business understanding* sebesar 10-20%, serta *deployment* sebesar 5-10% [9]. Fokus dalam penelitian ini adalah 3 tahap awal yaitu *business understanding*, *data understanding*, dan *data preparation*.



Gambar 1. *Metode Penelitian*

1) *Business Understanding*

Tahap ini menekankan pemahaman kebutuhan bisnis agar dapat diterjemahkan menjadi perumusan masalah data mining.

Pertama, dilakukan penentuan tujuan bisnis dengan mengenali kebutuhan utama organisasi, mengidentifikasi permasalahan, serta meng-klarifikasi harapan stakeholder [8], [9]. Kedua, penilaian situasi dilakukan melalui analisis kondisi, sumber daya, batasan, risiko, serta asumsi, yang biasanya disusun dalam bentuk analisis SWOT untuk memberikan gambaran menyeluruh [9]. Ketiga, disusun rencana proyek yang memuat daftar aktivitas, alokasi sumber daya, input-output, serta ketergantungan antar tahap [9].

2) *Data Understanding*

Tahap ini bertujuan mengenal karakteristik data lebih mendalam melalui empat sub-tahap. Pertama, *collect initial data* yaitu mengumpulkan data awal yang relevan dengan tujuan dan memuatnya ke dalam sistem analisis . Kedua, *describe data* untuk mengevaluasi karakteristik umum data melalui laporan deskriptif mengenai format, jumlah, struktur, serta identitas fitur [9]. Ketiga, *explore data* yaitu eksplorasi awal menggunakan visualisasi dan analisis sederhana untuk menemukan pola, anomali, atau struktur yang menarik [9]. Keempat, *verify data quality* dengan menguji kelengkapan dan kebenaran data, termasuk mengidentifikasi kesalahan isian serta frekuensinya [9].

3) *Data Preparation*

Tahap ini merupakan yang paling memakan waktu dan mencakup lima sub-tahap. Pertama, *select data* yaitu pemilihan data yang sesuai dengan tujuan analisis, mempertimbangkan kualitas serta keterbatasan teknis[9]. Kedua, *clean data* mencakup penanganan noise, *missing value*, dan *outlier* dengan metode statistik seperti *Interquartile Range* [10]. Ketiga, *integrate data* yaitu penggabungan informasi dari berbagai sumber atau tabel untuk memperoleh data yang lebih lengkap melalui teknik merging dan aggregation [9]. Keempat, *construct data* berupa pembuatan fitur turunan, transformasi nilai, serta normalisasi menggunakan metode seperti Min-Max, Z-Score, Box-Cox, atau Yeo-Johnson[9], [11]. Kelima, *format data* yang berfokus pada transformasi sintaksis agar sesuai dengan kebutuhan alat pemodelan tanpa mengubah makna data [9].

Berdasarkan SNI, kualitas air budidaya ikan air tawar secara spesifik untuk beberapa ikan seperti ikan nila, ikan mas, ikan patin, ikan gurami, dan ikan lele dengan wadah kolam dapat dilihat pada Tabel 1.

Tabel 1. *Atribut kualitar air berdasarkan jenis ikan*

Jenis Ikan	Suhu (C)	pH	DO (mg/l)	Amonia (mg/l)	Nitrat (mg/l)
Ikan Nila (Oreochromis niloticus Bleeker)	23-30	6,5 – 8,5	> 5	< 0,02	<50
Ikan Mas (Cyprinus carpio Linneaus)	25-30	6,5 – 8,5	> 5	< 0,01	<50
Ikan Patin (Pangasius) hypophthalmus	25-30	6,5 - 8,5	> 4	< 0,01	< 50
Ikan Gurami (Osphronemous gouramy Lac.)	25-30	6,5-8,6	>2	< 0,02	< 50
Ikan Lele (Clarias sp.)	25-30	6,5-8,5	> 4	< 0,01	<50

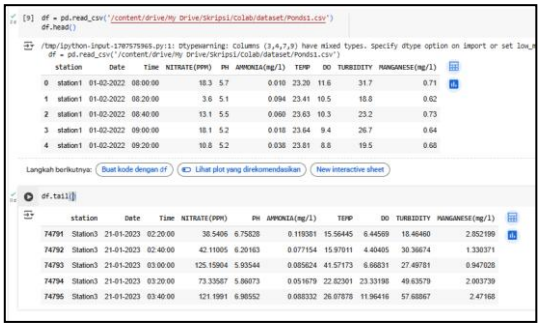
Untuk mengetahui kualitas air, sensor digunakan sebagai pengukur parameter kualitas air. Penggunaan sensor dalam pengambilan data kualitas air memberikan kemudahan dalam pemantauan secara real-time dan efisien [2]. Khusus atribut Amonia (NH_3), Menurut U.S. Environmental Protection Agency, 2025, amonia bebas (NH_3) tidak diukur langsung, melainkan dihitung dari total ammonia nitrogen (TAN), pH, dan suhu [12]. Hal ini sejalan dengan Francis-Floyd et al., 2023, yang menyatakan bahwa sebagian besar kit komersial hanya mengukur TAN dalam mg/L, bukan NH_3 [13]. Proporsi TAN yang bersifat toksik dapat meningkat dengan kenaikan suhu dan pH, sehingga perlu menentukan nilai NH_3 melalui TAN [14]. Penghitungan NH3 berdasarkan nilai TAN mengikut rumusan tertentu, namun dapat diperoleh melalui alat bantu contohnya Un-Ionized Ammonia Calculator (https://www.svl.net/ unionized-amonia-calculator/) [15].

HASIL DAN PEMBAHASAN

Hasil *data understanding*

Penelitian ini menggunakan dataset kualitas air yang diperoleh dari Kaggle dengan judul dataset Pondsdata (https://www.kaggle-.com/datasets/apgopi/pondsdata) dengan total 74.796 baris data dan 9

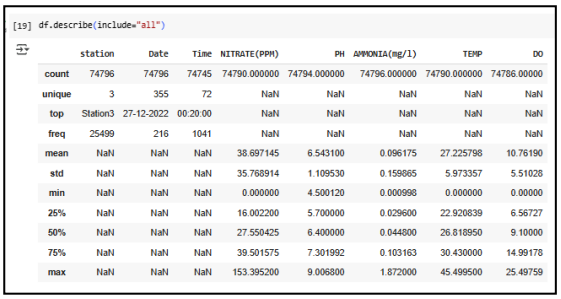
atribut, meliputi station, *Date*, *Time*, *NITRATE(PPM)*, *PH*, *AM-MONIA(mg/l)*, *TEMP*, *TURBIDITY*, dan *MANGANESE(mg/l)*. Data dikumpulkan secara real-time menggunakan teknologi *Internet of Things* (IoT) dengan sensor kualitas air yang terhubung melalui Arduino dan NodeMCU (ESP8266) di tiga kolam ikan berbeda di Guntur, Andhra Pradesh, India, selama periode satu tahun dari Februari 2022 hingga Januari 2023.



Gambar 2. *Hasil collect initial data*

Analisis statistik deskriptif menunjukkan bahwa setiap atribut memiliki jumlah baris yang berbeda, mengindikasikan adanya *missing value*. Atribut *station* dan *Date* memiliki jumlah baris terlengkap (74.796), sedangkan atribut lainnya memiliki jumlah yang lebih sedikit. Dataset menunjukkan karakteristik pengambilan data dari tiga stasiun dengan cakupan 355 hari dan pengambilan sampel setiap 20 menit, menghasilkan 72 waktu pengamatan per hari. Nilai rata-rata untuk atribut kualitas air menunjukkan *NITRATE(PPM)* sebesar 38,70, PH sebesar 6,54, AMMONIA(mg/l) sebesar 0,096, TEMP sebesar 27,23°C, dan DO sebesar 10,76 mg/l.

Analisis distribusi data mengungkapkan bahwa *NITRATE(PPM)* dan *AMMONIA(mg/l)* menunjukkan distribusi *right-skewed* dengan keberadaan *outlier* yang signifikan, terutama nilai *NITRATE* di atas 50 PPM dan *AMMONIA* di atas 0,25 mg/l.

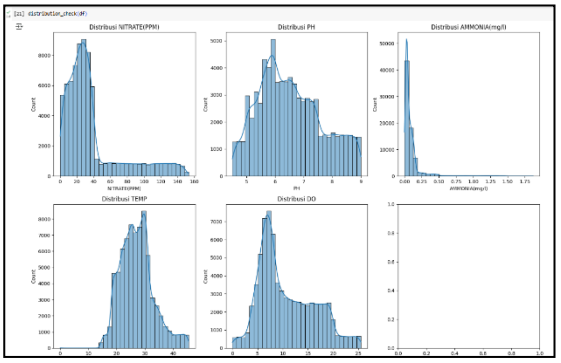


[19] df.describe(include="all")

	station	Date	Time	NITRATE(PPM)	PH	AMMONIA(mg/l)	TEMP	DO
count	74796	74796	74745	74790.000000	74794.000000	74796.000000	74790.000000	74786.00000
unique	3	355	72	NaN	NaN	NaN	NaN	NaN
top	Station3	27-12-2022	00:20:00	NaN	NaN	NaN	NaN	NaN
freq	25499	216	1041	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	38.697145	6.543100	0.096175	27.225798	10.76190
std	NaN	NaN	NaN	35.768914	1.109530	0.159865	5.973357	5.51028
min	NaN	NaN	NaN	0.000000	4.500120	0.000998	0.000000	0.00000
25%	NaN	NaN	NaN	16.002200	5.700000	0.029600	22.920839	6.56727
50%	NaN	NaN	NaN	27.550425	6.400000	0.044800	26.818950	9.10000
75%	NaN	NaN	NaN	39.501575	7.301992	0.103163	30.430000	14.99178
max	NaN	NaN	NaN	153.395200	9.006800	1.872000	45.499500	25.49759

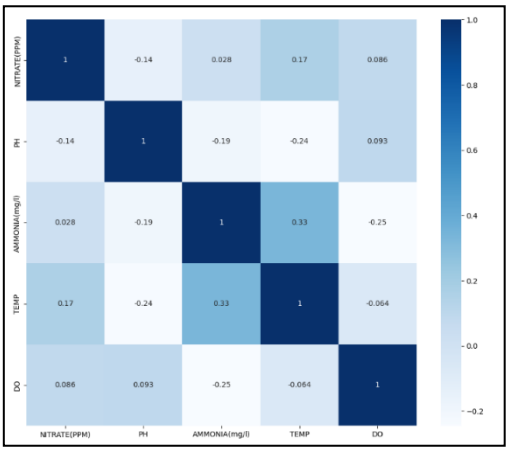
Gambar 3. *Hasil describe data*

Atribut *PH* menunjukkan distribusi mendekati normal dengan sebagian besar nilai berada di rentang 6-7, sementara *TEMP* menunjukkan distribusi normal namun terdapat nilai *outlier* mendekati 0°C. Atribut *DO* menunjukkan distribusi sedikit *right-skewed* dengan dominasi nilai di sekitar 6-12 mg/l.



Gambar 4. *Hasil distribusi data*

Analisis korelasi menunjukkan hubungan positif terkuat antara *TEMP* dan *AMMONIA(mg/l)* dengan nilai 0,33, sedangkan hubungan negatif terkuat terlihat antara *PH* dan *TEMP* (-0,24) serta *AMMONIA(mg/l)* dan *DO* (-0,25).



Gambar 5. *Korelasi heatmap*

Hasil *Data Preparation*

Tahap *data preparation* dimulai dengan seleksi atribut yang relevan, yaitu *NITRATE(PPM)*, *PH*, *AMMONIA(mg/l)*, *TEMP*, dan *DO*, sementara atribut *TURBIDITY* dan *MANGANESE(mg/l)* dihapus untuk mengurangi penggunaan memori dan meningkatkan efisiensi komputasi. *Missing value* ditangani melalui imputasi menggunakan median untuk data numerik dan modus untuk data non-numerik, dengan hasil bahwa semua missing value berhasil diimputasi hingga mencapai nilai 0.

```
{'validation': {'rows_remaining': 74796,
  'missing_values_remaining': np.int64(0),
  'duplicates_remaining': np.int64(0),
  'data_loss_percentage': 0.0}}
```

Gambar 6. *Hasil validasi data*

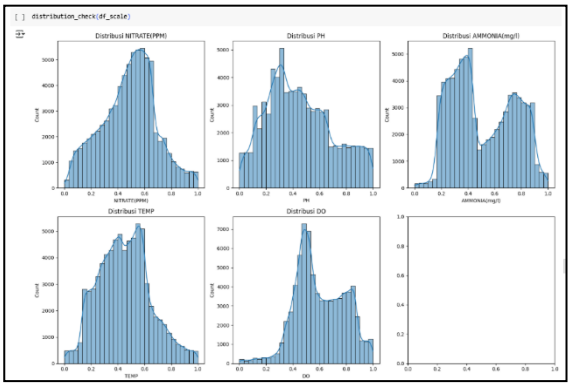
Penanganan *outlier* dilakukan pada tiga atribut utama yang teridentifikasi memiliki outlier signifikan, yaitu NITRATE(PPM), AMMONIA(mg/l), dan TEMP. Metode imputasi menggunakan median dipilih karena distribusi data yang tidak normal dan cenderung *right-skewed*. Setelah imputasi outlier, terjadi perubahan signifikan pada kuartil ketiga atribut tersebut. NITRATE(PPM) mengalami penurunan nilai maksimum dari 153,40 menjadi 74,74, AMMONIA(mg/l) dari 1,87 menjadi 0,21 mg/l, dan TEMP mengalami perbaikan distribusi dengan nilai minimum meningkat dari 0 menjadi 12,50°C dan maksimum menurun dari 45,50 menjadi 41,69°C.

```
# untuk melihat statistik data
df_imputed.describe(include="all")
```

	NITRATE(PPM)	AMMONIA(mg/l)	TEMP	DO	PH	station	Date	Time
count	74796.000000	74796.000000	74796.000000	74796.000000	74796.000000	74796	74796	74796
unique	NaN	NaN	NaN	NaN	NaN	3	355	72
top	NaN	NaN	NaN	NaN	NaN	Station3	27-12-2022	00:00:00
freq	NaN	NaN	NaN	NaN	NaN	25499	216	1092
mean	27.157757	0.064000	26.786165	10.761677	6.543096	NaN	NaN	NaN
std	15.381043	0.042456	5.306428	5.509945	1.109516	NaN	NaN	NaN
min	0.000000	0.000998	14.500380	0.000000	4.500120	NaN	NaN	NaN
25%	15.729050	0.029600	22.889960	6.567270	5.700000	NaN	NaN	NaN
50%	26.757850	0.044800	26.635891	9.100000	6.400000	NaN	NaN	NaN
75%	36.422975	0.095770	30.154625	14.991105	7.301978	NaN	NaN	NaN
max	74.738250	0.213185	41.688700	25.497590	9.006800	NaN	NaN	NaN

Gambar 7. *Hasil describe akhir*

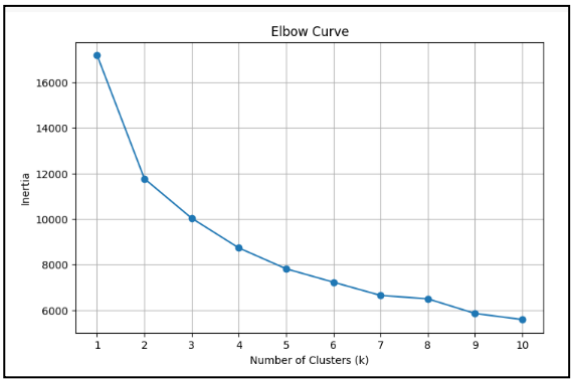
Tahap *construct data* menghasilkan transformasi distribusi yang signifikan pada semua atribut. *NITRATE(PPM)*, *TEMP*, dan *PH* menunjukkan distribusi yang hampir mendekati normal, sementara *AMMONIA(mg/l)* dan *DO* membentuk distribusi bimodal dengan dua puncak, mengindikasikan adanya dua kelompok berbeda dalam atribut tersebut. Nilai *skewness* semua atribut menunjukkan perbaikan mendekati distribusi normal, dan analisis korelasi menunjukkan hubungan lemah antara *PH* dengan *NITRATE(PPM)* serta *TEMP* dengan *AMMONIA(mg/l)*.



Gambar 8. *Distribusi data*

Hasil *Clustering*

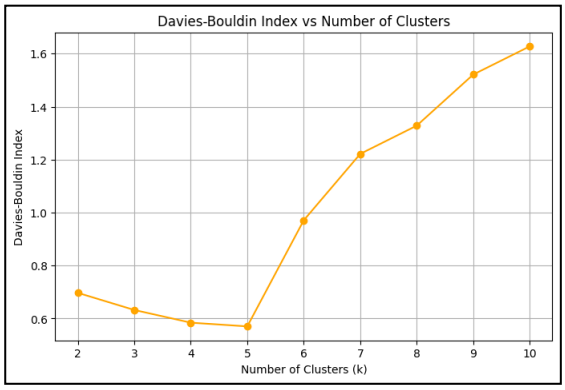
Implementasi algoritma *K-Means clustering* dimulai dengan penentuan nilai K optimal menggunakan metode *Elbow*, yang menunjukkan bahwa K=5 merupakan nilai optimal untuk dataset ini. Penerapan algoritma *K-Means* menghasilkan lima klaster yang masing-masing merepresentasikan kelompok data dengan karakteristik kualitas air yang berbeda. Hasil *clustering* kemudian ditransformasi kembali ke bentuk data asli untuk memudahkan interpretasi.



Gambar 9. *Hasil Metode Elbow*

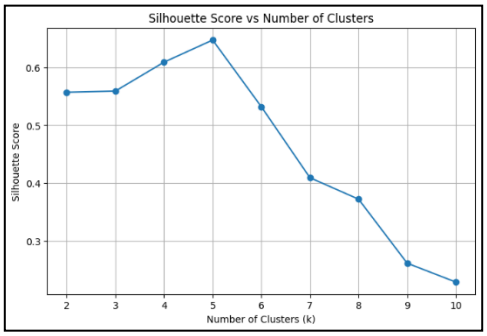
Grafik *Davies-Bouldin Index* (DBI) memperlihatkan tren penurunan nilai hingga mencapai titik terendah pada K=5, kemudian kembali meningkat pada jumlah klaster

berikutnya. Pola tersebut menunjukkan bahwa hingga K=5, kualitas pemisahan antar klaster semakin baik karena jarak antar klaster semakin besar dan keseragaman dalam klaster semakin tinggi. Namun, setelah melewati titik tersebut, penambahan jumlah klaster justru menurunkan kualitas karena beberapa klaster menjadi terlalu berdekatan atau tumpang tindih.



Gambar 10. *Hasil Metode Davies-Bouldin Index*

Pada grafik *Silhouette Coefficient*, terlihat bahwa nilai koefisien meningkat seiring bertambahnya jumlah klaster hingga mencapai puncaknya pada K=5, lalu menurun kembali setelahnya. Pola ini menunjukkan bahwa pada K=5, pemisahan antar klaster paling optimal dan objek dalam satu klaster memiliki kemiripan yang tinggi. Penurunan setelah titik tersebut mengindikasikan bahwa pembagian menjadi lebih banyak klaster justru membuat batas antar kelompok menjadi kurang jelas.



Gambar 11. *Hasil Metode Silhouette Score*

Evaluasi hasil pengelompokan menggunakan algoritma *K-Means* dilakukan dengan mempertimbangkan tiga metrik, yaitu WCSS (*Inertia*), *Silhouette Score*, dan *Davies-Bouldin Index* (DBI). Berdasarkan gambar hasil pengujian, nilai WCSS menunjukkan penurunan signifikan saat jumlah klaster meningkat dari 2 hingga 5, kemudian mulai melandai pada K>5, menandakan bahwa tambahan klaster setelah titik tersebut tidak memberikan peningkatan signifikan terhadap kualitas *clustering*. *Silhouette Score* meningkat hingga puncaknya pada K=5 sebelum menurun kembali pada jumlah klaster lebih besar, yang menunjukkan bahwa pemisahan antar klaster dan keseragaman data dalam klaster paling optimal pada titik ini. Sementara itu, DBI mencapai nilai terendah pada K=5, menandakan jarak antar klaster terbesar dan penyebaran dalam klaster terkecil. Berdasarkan ketiga metrik ini, jumlah klaster optimal ditetapkan pada K=5, karena memberikan keseimbangan terbaik antara kedekatan dalam klaster dan pemisahan antar klaster.

	k	WCSS (Inertia)	Silhouette Score	Davies-Bouldin Index
0	2	11759.648897	0.287486	1.406637
1	3	10037.696117	0.227714	1.494669
2	4	8714.413205	0.239044	1.539589
3	5	7811.776394	0.251104	1.380780
4	6	7379.891510	0.221496	1.484138
5	7	7020.962213	0.205316	1.645840
6	8	6212.999579	0.216543	1.418428
7	9	5878.617124	0.198285	1.516168
8	10	5671.139633	0.198576	1.538342

Gambar 12. *Hasil pengujian WCSS, Silhouette Score, dan Davies Bouldin Index*

Setiap klaster yang terbentuk mencerminkan kondisi kualitas air yang spesifik dengan kombinasi atribut yang berbeda, memberikan insight berharga untuk strategi pengelolaan dan monitoring sistem akuakultur.

cluster	NITRATE(PPM)	AMMONIA(mg/l)	TEMP	DO	PH
0	25.672109	0.107656	31.017476	6.674666	5.874596
1	20.319246	0.030344	24.400985	12.749480	8.002215
2	18.808906	0.033013	25.052720	11.217983	6.056002
3	44.877043	0.099948	27.973423	18.100711	5.951250
4	46.729337	0.056835	23.620314	4.370857	6.915975

Gambar 13. *Hasil clustering*

Dengan menggunakan *Un-Ionized Ammonia Calculator,* amonia dikonversi dari TAN menjadi NH_3. Kemungkinan sensor yang digunakan lebih mengukur *Total Ammonia Nitrogen* (TAN) daripada amonia bebas (NH_3).

Tabel 2 adalah hasil konversi menggunakan kalkulator *Un-Ionized Ammonia Calculator.*

Berdasarkan standar SNI dan toleransi [7], hanya klaster 1 yang konsisten memenuhi atribut optimal bagi budidaya ikan nila, mas, patin, gurami, dan lele. Klaster 2 masih dapat ditoleransi meski tidak seideal klaster 1, sedangkan klaster 0, 3, dan 4 menunjukkan kondisi kurang ideal akibat pH terlalu asam atau DO rendah. Dengan demikian, klaster 1 dan 2 paling sesuai untuk budidaya, kecuali klaster 4 yang masih dapat ditoleransi oleh ikan lele.

Tabel 2. *Hasil konversi TAN ke NH_3*

Klaster	TAN	NH_3	Persentasi NH_3
Klaster 0	0.107656	0,000	0,1%
Klaster 1	0.107656	0,002	5,2%
Klaster 2	0.107656	0,000	0,1%
Klaster 3	0.099948	0,000	0,1%
Klaster 4	0.056835	0,000	0,4%

Hasil clustering dibandingkan dengan standar kualitas air berdasarkan SNI untuk tiga jenis ikan air tawar, yaitu ikan nila (Oreochromis niloticus), ikan mas (Cyprinus carpio), ikan patin (Pangasius hypophthalmusikan), gurami (Osphronemous gouramy Lac.), dan ikan lele (Clarias sp.). Ikan nila membutuhkan suhu 23–30°C, pH 6,5–8,5, DO > 5 mg/L, amonia (NH_3) <0,02 mg/L, dan nitrat <50 mg/L. Dari hasil clustering, suhu dan nitrat pada seluruh klaster sesuai dengan standar. DO sesuai pada klaster 0, 1, 2, dan 3, namun tidak pada klaster 4. Untuk pH, hanya klaster 1 (8,00) dan klaster 4 (6,91) yang sesuai. Amonia (NH_3) sesuai pada semua klaster.

Ikan mas membutuhkan suhu 25–30°C, pH 6,5–8,5, DO > 5 mg/L, amonia <0,01 mg/L, dan nitrat <50 mg/L. Hasil analisis menunjukkan bahwa suhu sesuai pada klaster 0, 2, dan 3, tetapi klaster 1 (24,40°C) dan klaster 4 (23,62°C) berada di bawah SNI. Parameter DO memenuhi syarat pada klaster 0, 1, 2, dan 3, namun tidak pada klaster 4. pH sesuai hanya pada klaster 1 dan 4. Meskipun begitu, kadar amonia (NH_3) seluruh klaster melebihi ambang batas (<0,01 mg/L). Amonia (NH_3) sesuai pada semua klaster.

Ikan patin membutuhkan suhu 25–30°C, pH 6,5–8,5, DO > 4 mg/L, amonia <0,01 mg/L, dan nitrat <50 mg/L. Berdasarkan hasil clustering, suhu sesuai pada klaster 0, 2, dan 3, sedangkan klaster 1 (24,40°C) dan klaster 4 (23,62°C) berada di bawah standar minimal. Parameter DO memenuhi syarat pada seluruh klaster karena nilainya di atas 4 mg/L, meskipun klaster 4 berada pada ambang batas (4,37 mg/L). pH sesuai pada klaster 1 dan 4, sementara klaster lainnya tidak memenuhi. Amonia (NH_3) sesuai pada semua klaster. Parameter nitrat juga sesuai karena seluruh klaster berada di bawah 50 mg/L.

Ikan gurami membutuhkan suhu 25–30°C, pH 6,5–8,6, DO > 2 mg/L, amonia <0,02 mg/L, dan nitrat <50 mg/L. Hasil clustering menunjukkan bahwa suhu sesuai pada klaster 0, 2, dan 3, tetapi klaster 1 (24,40°C) dan klaster 4 (23,62°C) berada di bawah batas minimal. Parameter DO memenuhi syarat pada semua klaster karena nilainya lebih dari 2 mg/L, meskipun klaster 4 berada pada nilai terendah (4,37 mg/L) namun masih aman. pH sesuai pada klaster 1 dan 4, sedangkan klaster lain tidak memenuhi. Amonia (NH_3) sudah sesuai pada semua klaster. Sementara itu, nitrat juga sesuai karena nilainya masih di bawah 50 mg/L di semua klaster.

Ikan lele membutuhkan suhu 25–30°C, pH 6,5–8,5, DO > 4 mg/L, amonia <0,01 mg/L, dan nitrat <50 mg/L. Hasil clustering memperlihatkan bahwa suhu sesuai pada klaster 0, 2, dan 3, tetapi klaster 1 dan 4 berada di bawah batas minimal. Parameter DO sesuai pada semua klaster karena nilainya berada di atas 4 mg/L, meskipun klaster 4 berada pada ambang batas (4,37 mg/L). pH sesuai pada klaster 1 dan 4. Amonia (NH_3) sesuai pada semua klaster.

Table 2. *Klaster yang cocok dengan ikan*

Jenis Ikan	Klaster				
	0	1	2	3	4
Ikan Nila		✓	✓		
Ikan Mas		✓	✓		
Ikan Patin		✓	✓		
Ikan Gurami		✓	✓		
Ikan Lele		✓	✓	✓	

Jika berdasarkan standar SNI dan teori toleransi Boyd (1998), hanya klaster 1 yang secara konsisten memenuhi semua atribut optimal untuk budidaya ikan nila, ikan mas, ikan patin, ikan gurami dan lele. Klaster 2 masih dapat ditoleransi meskipun tidak seideal klaster 1. Sementara itu, klaster 0, 3, dan 4 menunjukkan kondisi yang tidak ideal akibat atribut seperti pH yang terlalu asam (klaster 0 dengan pH 5,87 dan klaster 3 dengan pH 5,95), dan kadar DO yang cukup rendah (klaster 4 dengan DO 4,37 mg/L). Dengan demikian, klaster 1 dan 2 merupakan kelompok yang paling konsisten memenuhi standar kualitas air bagi ketiga jenis ikan. Kecuali klaster 4 yang masih dapat ditoleransi oleh ikan Lele.

Jika menggunakan K = 2, didapat 2 klaster yaitu klaster 0 dan klaster 1. Klaster 0 memiliki karakteristik kualitas air dengan nilai pH yang lebih rendah yaitu 5.91, menandakan kondisi air yang cenderung asam. Selain itu, klaster ini juga menunjukkan nilai amonia (NH_3) dan nitrat yang lebih tinggi, serta suhu air yang lebih hangat yaitu 29.19°C. DO pada klaster ini relatif lebih rendah dibanding klaster lainnya. Sebaliknya, klaster 1 ditandai dengan pH yang lebih netral yaitu 7.17, serta konsentrasi ammonia (NH_3) dan nitrat yang lebih rendah, yang menunjukkan kualitas air yang lebih baik. Klaster ini juga memiliki suhu yang lebih. rendah yaitu 24.44°C dan oksigen terlarut yang lebih tinggi yaitu 11.75 mg/l. Klaster 1 merupakan klaster yang ideal untuk ikan Nila, ikan Mas, dan ikan Lele. Hasil dapat dilihat pada Gambar 11.

cluster	NITRATE(PPM)	AMMONIA(mg/l)	TEMP	DO	PH
0	32.848551	0.097772	29.190698	9.752531	5.906201
1	21.596341	0.030995	24.436299	11.747881	7.165512

Gambar 14. *Hasil statistik deskriptif clustering dengan K = 2*

PENUTUP

Penelitian ini menunjukkan bahwa kualitas data sensor air memerlukan tahapan *preprocessing* untuk mengatasi ketidak-sesuaian, *outlier*, dan *missing value* agar hasil analisis lebih akurat. Analisis deskriptif terhadap 74.796 data pengukuran mengungkap variasi cukup besar pada setiap atribut, dengan suhu rata-rata 26,79°C, pH 6,54, DO 10,76 mg/L, nitrat 27,16 ppm, dan amonia 0,0640 mg/L. Terdapat hubungan positif antara suhu dan amonia, serta hubungan negatif antara suhu dan pH, serta amonia dan DO. Hasil *clustering* menghasilkan lima klaster dengan karakteristik kualitas air berbeda. Berdasarkan SNI dan toleransi Boyd, klaster 1 merupakan kondisi ideal bagi ikan nila, mas, patin, gurami, dan lele, sementara klaster 2 masih dapat ditoleransi meski kurang optimal. Sebaliknya, klaster 0, 3, dan 4 tidak sesuai karena pH asam dan DO rendah. Selain itu, perlu dicatat bahwa sensor kemungkinan mengukur *Total Ammonia Nitrogen* (TAN), sehingga dilakukan konversi ke amonia bebas (NH_3) menggunakan *Un-Ionized Ammonia Calculator*. Ke depan, hasil penelitian ini dapat dikembangkan lebih lanjut dengan membangun model prediksi atau sistem rekomendasi spesies ikan berbasis AIoT yang memanfaatkan hasil *preprocessing* dan *clustering* sebagai fondasi utama.

DAFTAR PUSTAKA

[1] M. Mugwanya, M. A. O. Dawood, F. Kimera, and H. Sewilam, "A review on recirculating aquaculture system: influence of stocking density on fish and crustacean behavior, growth performance, and immunity," *Ann. Anim. Sci.*, vol. 22, no. 3, pp. 873–884, July 2022, doi: 10.2478/aoas-2022-0014.

[2] S. N. Zainurin *et al.*, "Advancements in Monitoring Water Quality Based on Various Sensing Methods: A Systematic Review," *Int. J. Environ. Res. Public. Health*, vol. 19, no. 21, p. 14080, Oct. 2022, doi: 10.3390/ijerph192114080.

[3] A. Khumaidi, "DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE PRODUCTION USING

CRISP-DM METHOD," *J. Techno Nusa Mandiri*, vol. 17, no. 1, pp. 1–8, Feb. 2020, doi: 10.33480/techno.v17i1.1240.

[4] F. Dwiatmoko, E. Utami, and Sudarmawan, "Preprocessing Tranformasi Data Menggunakan K-Means Clustering," *J. EXPLORE*, vol. 11, no. 2, 2021, doi: https://doi.org/10.35200/explore.v11i2.544.

[5] R. A. Sigit, V. R. Akbar, Rahmaddeni, S. Aziz, and E. R. Pebriansyah, "Penerapan Algoritma K-Means Clustering dalam Menganalisis Pola Peminjaman Buku di Perpustakaan," *Indones. J. Comput. Sci.*, vol. 13, no. 5, Oct. 2024, doi: 10.33022/ijcs.v13i5.4317.

[6] D. Setiadi, B. Irawan, and A. Bahtiar, "PENERAPAN ALGORITMA K-MEANS CLUSTERING PADA PEMBESARAN JENIS IKAN UNGGULAN DI PROVINSI JAWA BARAT," *J. Mhs. Tek. Inform.*, vol. 7, no. 6, 2023, doi: https://doi.org/10.36040/jati.v7i6.8179.

[7] C. E. Boyd, "Water Quality for Pond Aquaculture," Alabama Agricultural Experiment Station, Auburn University, Auburn, Alabama, Aug. 1998.

[8] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks," *SBC Rev. Comput. Sci.*, vol. 4, no. 1, pp. 28–43, Oct. 2024, doi: 10.5753/reviews.2024.3757.

[9] P. Chapman *et al.*, *CRISP-DM 1.0: Step-by-step Data Mining Guide*. 2000.

[10] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. in Intelligent Systems Reference Library. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-10247-4.

[11] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale Univ. J. Eng. Sci.*, vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.

[12] U.S. Environmental Protection Agency, "Ammonia," CADDIS – Causal Analysis/Diagnosis Decision Information System. Accessed: Sept. 05, 2025. [Online]. Available: https://www.epa.gov/ caddis/ammonia

[13] R. Francis-Floyd, C. Watson, D. Petty, and D. B. Pouder, "Ammonia in Aquatic Systems," *UFIFAS EDIS Univ. Fla. George Smathers Libr.*, Dec. 2023, doi: https://doi.org/10.32473/edis-fa031-2022.

[14] R. M. Durborow, D. M. Crosby, and M. W. Brunson, "Ammonia in fish ponds (SRAC Publication No. 463)," Cooperative Extension Program, Kentucky State University; Virginia State University; Mississippi Cooperative Extension Service, 2019. [Online]. Available: https://aquaculture.ca.uky.edu/sites/aquaculture.ca.uky.edu/files/srac_463_ammonia_in_fish_ponds.pdf

[15] L. O. M. H. Akbar, A. Gappar, and D. A. Christyandari, "Konversi kadar amonia (NH_3) dari amonia total (NH_3-N) menggunakan alat bantu konversi tanpa data salinitas," *J. Bul. Tek. Litkayasa Akuakultur*, vol. 17, no. 2, pp. 161–165, 2019, doi: https://doi.org/10.15578/blta.17.2.2019.161-165.

Halaman ini sengaja dikosongkan.