SAINTIFIK:
Jurnal Matematika, Sains, dan Pembelajarannya **Vol.11, No.2, Juli 2025, pp. 285~295**
ISSN 2407-4098 (print) DOI:10.31605/saintifik.v11i2.611
ISSN 2622-8904 (online)

Klasifikasi Kualitas Air Sungai dengan Metode Random Forest

Muhammad Afrizal Tanjung [*1], Rima Aprilia[2]
[1,2]Program Studi Matematika, Universitas Islam Negeri Sumatera Utara, Indonesia
e-mail: muhammad0703192035@uinsu.ac.id

Abstrak

Kualitas air sungai memegang peranan penting bagi kesehatan publik dan pelayanan perkotaan, namun banyak sungai di Indonesia menunjukkan indikasi pencemaran. Studi ini menerapkan algoritma Random Forest untuk mengklasifikasikan mutu air tiga sungai di Kota Medan berdasarkan data pemantauan sekunder tahun 2023–2024 dari Dinas Lingkungan Hidup. Dataset berisi 72 observasi dengan sembilan parameter utama, yaitu TSS, pH, BOD, COD, DO, Nitrat, Nitrit, Total Coliform, dan Amonia. Skema pemodelan meliputi pra pengolahan data, pembagian latih–uji 80:20 secara terstratifikasi, pelatihan Random Forest dengan 100 pohon, serta evaluasi menggunakan akurasi dan matriks kebingungan pada subset uji. Hasil menunjukkan akurasi keseluruhan 100 persen pada data uji, dengan ketepatan penuh pada kedua kelas yang dikaji (Kelas II dan Kelas III). Analisis kepentingan fitur mengindikasikan bahwa Total Coliform dan COD merupakan penentu paling dominan, diikuti Nitrat dan DO, sedangkan TSS, pH, Ammonia, dan parameter lain memberi kontribusi menengah hingga rendah. Temuan ini menegaskan efektivitas Random Forest untuk tugas klasifikasi mutu air sungai dan memberikan wawasan prioritas parameter bagi pengendalian pencemaran. Secara praktis, pendekatan ini dapat mendukung pemantauan berbasis data dan pengambilan keputusan pengelolaan kualitas air di tingkat daerah.

Kata kunci*— Kualitas Air Sungai, Random Forest, Klasifikasi, Kota Medan, RapidMiner.*

1. PENDAHULUAN

Kebutuhan air bersih untuk konsumsi rumah tangga, aktivitas ekonomi, dan layanan publik terus meningkat seiring pertumbuhan penduduk dan urbanisasi (Alihar, 2018; Suryani, 2020; Febriawati *et al.,* 2021; Januari *et al.,* 2024). Secara hidrologis, sumber air terbagi menjadi air hujan, air tanah, dan air permukaan; sungai merupakan salah satu sumber yang paling mudah dijumpai dan dimanfaatkan masyarakat (Nugroho *et al.,* 2020; Sukristiyono *et al.,* 2021). Air bersih idealnya tidak berwarna, tidak berbau, tidak berasa, dan memenuhi persyaratan kualitas sesuai peruntukan.

Berbagai laporan nasional menunjukkan kualitas air sungai di Indonesia masih memprihatinkan. Data Direktorat Jenderal Pengendalian Pencemaran Kementerian Lingkungan Hidup dan Kehutanan pada 2015 mengindikasikan porsi signifikan sungai berada pada kondisi tercemar, sementara pemantauan Badan Pusat Statistik tahun 2017 pada puluhan sungai di 34 provinsi juga menunjukkan prevalensi pencemaran. Sumber pencemar berasal dari limbah domestik, aktivitas industri dan pertambangan, serta transportasi air (Rosyidah, 2018; Pratiwi, 2020; Liku *et al.,* 2022; Indriyani *et al.,* 2024). Secara regulatif, Peraturan Pemerintah Republik Indonesia No. 82 Tahun 2001 mendefinisikan pencemaran air sebagai masuknya makhluk hidup, zat, energi, dan/atau komponen lain ke badan air oleh kegiatan manusia sehingga kualitas air menurun hingga tidak berfungsi sesuai peruntukannya.

Di kawasan perkotaan padat seperti Kota Medan yang kerap mengalami banjir dan tekanan kualitas lingkungan degradasi kualitas air sungai berpotensi mempengaruhi ketersediaan air bersih dan berdampak pada kesehatan publik. Kondisi ini menuntut strategi penilaian mutu air yang cepat, andal, dan scalable agar pemangku kepentingan dapat merencanakan intervensi secara tepat sasaran. Penilaian kualitas air umumnya mengandalkan pengukuran parameter fisika–kimia seperti pH, DO, BOD, COD, TSS, Fe, dan Mn, serta perbandingan terhadap baku mutu; namun pendekatan konvensional menghadapi tantangan ketika data berukuran besar, memiliki hubungan non linier antar parameter, dan mengandung noise (Wijaya *et al.,* 2024).

Metode pembelajaran mesin menawarkan solusi untuk tantangan tersebut, khususnya pada tugas klasifikasi berbasis banyak fitur. Random Forest pengembangan dari Classification and Regression Tree

(CART) dengan teknik bootstrap aggregating (bagging) dan random feature selection dikenal memiliki akurasi tinggi, robust terhadap overfitting, minim pra proses, serta menyediakan estimasi tingkat kepentingan variabel prediktor (feature importance) (SP *et al.,* 2023; Fauziah, 2025). Karakteristik ini relevan bagi data kualitas air yang multivariat dan heterogen.

Artikel ini mengajukan penerapan Random Forest untuk mengklasifikasikan kualitas air sungai di Kota Medan dengan studi kasus pada tiga sungai yang dipantau oleh Badan/Dinas Lingkungan Hidup Kota Medan tahun 2025. Keluaran model dirancang dalam dua kelas, yaitu layak dan tidak layak, yang diturunkan dari konsolidasi kriteria baku mutu sesuai peruntukan. Selain mengevaluasi kinerja model melalui ukuran akurasi, studi ini juga menganalisis kontribusi setiap parameter terhadap keputusan klasifikasi melalui feature importance guna memberikan wawasan praktis bagi pengelolaan kualitas air.

Kontribusi utama penelitian ini mencakup perumusan kerangka kerja klasifikasi mutu air sungai berbasis Random Forest pada konteks perkotaan Indonesia, penyajian evaluasi kinerja model dua kelas yang dapat dijadikan dasar pemilihan model operasional, serta pengungkapan parameter paling berpengaruh untuk mendukung prioritas upaya pengendalian pencemaran di tingkat lokasi dan parameter. Pendekatan yang diusulkan diharapkan memperkaya praktikum pemantauan kualitas air melalui metode berbasis data yang transparan, dapat bereplikasi, dan berguna bagi perumusan kebijakan pengelolaan kualitas air yang berbasis bukti.

2. METODE PENELITIAN

Penelitian ini dilaksanakan di Dinas Lingkungan Hidup (DLH) Kota Medan yang beralamat di Jalan Pinang Baris No. 114, Medan, mulai Oktober 2025 hingga selesai, meliputi tahap perencanaan, pengumpulan data sekunder, pengolahan, pemodelan, evaluasi, dan penyusunan naskah ilmiah. Penelitian bersifat kuantitatif dengan memanfaatkan data sekunder resmi dari DLH Kota Medan berupa hasil pemantauan kualitas air sungai yang terdokumentasi pada dua lembar kerja untuk periode 2023 dan 2024. Unit analisis adalah setiap hasil pengukuran parameter kualitas air pada titik pantau di tiga sungai di Kota Medan, yakni Sungai Batuan, Sungai Belawan, dan Sungai Kera.

Variabel respon pada penelitian ini adalah status kelayakan kualitas air sungai dengan dua kategori, yaitu layak dan tidak layak sesuai peruntukan. Label kelas diturunkan dari penilaian kelas mutu I–IV menjadi dua kelas operasional (misalnya layak apabila memenuhi kriteria kelas I–II, dan tidak layak apabila termasuk kelas III–IV) atau mengikuti penandaan pada data DLH bila tersedia. Variabel prediktor merupakan parameter fisika-kimia dan mikrobiologi hasil uji laboratorium, meliputi TSS (mg/L), pH, BOD (mg/L), COD (mg/L), DO (mg/L), Nitrat $NO3^-$ (mg/L), Nitrit $NO2^-$ (mg/L), Total Coliform (MPN/100 mL), serta Amonia NH3/NH4+ (mg/L). Secara ringkas, TSS menggambarkan beban padatan tersuspensi yang memengaruhi kekeruhan; pH menunjukkan keasaman/kebasaan; BOD dan COD mengindikasikan beban bahan organik dan oksidan kimia; DO menunjukkan ketersediaan oksigen terlarut bagi biota; Nitrat dan Nitrit berkaitan dengan beban nutrien dan potensi eutrofikasi; Total Coliform menjadi indikator kontaminasi fekal; sedangkan Amonia mencerminkan beban organik dan potensi toksisitas bagi organisme air.

Data yang digunakan merupakan data sekunder dari DLH Kota Medan dalam format lembar kerja yang memuat identitas sungai dan titik pantau, waktu pengambilan sampel, serta nilai parameter yang diukur. Sebelum pengolahan, dilakukan verifikasi struktur data (nama kolom, satuan, dan format tanggal) serta pemeriksaan konsistensi dan kelengkapan. Nilai pengukuran kemudian dibandingkan dengan acuan baku mutu yang relevan untuk memperoleh label kelayakan sesuai skema klasifikasi dua kelas.

Adapun prosedur penelitian adalah sebagai berikut:

1. Perancangan studi (perumusan masalah, tujuan, indikator kinerja, penetapan variabel dan skema label, serta rencana analisis);
2. Akuisisi dan integrasi data (pengumpulan berkas tahun 2023–2024, penyeragaman skema kolom, dan penggabungan dataset);
3. Prapengolahan (pemeriksaan dan penanganan nilai hilang/duplikasi, penyelarasan satuan, deteksi pencilan berbasis domain dan statistik, serta pembuatan variabel turunan bila relevan);
4. Penetapan peran variabel dan penyandian label biner;
5. Pembagian data latih–uji secara stratified atau berbasis waktu bila runtun;

6. Penanganan ketidakseimbangan kelas pada data latih saja (misalnya oversampling terkontrol atau SMOTE);
7. Pelatihan model Random Forest;
8. Penalaan hyperparameter dengan validasi silang;
9. Evaluasi performa dengan akurasi, precision, recall, F1-score, AUC-ROC, dan matriks kebingungan;
10. Interpretasi hasil melalui feature importance, validasi, dan penyusunan laporan ilmiah.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data

Data yang digunakan dalam penelitian ini merupakan data pengukuran kualitas air sungai yang terdokumentasi dalam dua lembar kerja tahun 2023 dan 2024, masing-masing berisi 36 observasi sehingga total terdapat 72 baris data. Setiap observasi memuat identitas pengambilan sampel berupa variabel sungai, lokasi, dan bulan, serta parameter kualitas air utama yang meliputi TSS (mg/L), pH (tanpa satuan), BOD (mg/L), COD (mg/L), DO atau oksigen terlarut (mg/L), Nitrat (mg/L), dan Total Coliform (MPN/100 mL). Parameter-parameter tersebut mewakili aspek fisik, kimia, dan mikrobiologis yang lazim dipakai untuk menilai kondisi mutu perairan sungai. Data ini selanjutnya disiapkan untuk pemodelan klasifikasi kualitas air dengan algoritma Random Forest di RapidMiner, dengan peran variabel sungai, lokasi, dan bulan sebagai konteks spasial-temporal serta parameter numerik sebagai prediktor utama. Penelitian ini dilaksanakan di Dinas Lingkungan Hidup Kota Medan yang beralamat di Jalan Pinang Baris No.114.

Tabel 1. Deskripsi Data

SUNGAI	KRITERIA	2023		2024	
		Minimum	Maksimum	Minimum	Maksimum
Sei Batuan	TSS	2,1	38	1	22
	pH	6,13	7,45	6,43	7,54
	BOD	2,09	2,9	2,14	2,93
	COD	21,8	24,8	21,3	24,8
	DO	4	7,44	4,51	6,26
	Nitrat	1,7	3,11	0,48	1,67
	Nitrit	0,0097	0,18	0,0097	0,06
	Total Coliform	1700	5400	1300	4900
	Amonia	0,01	4,69	0,01	3,88
Sei Belawan	TSS	1	32	1	15
	pH	6,66	8,18	7,02	8,12
	BOD	2,52	2,9	1,43	2,87
	COD	21,2	24,8	20,8	23,6
	DO	4,3	7,36	5,43	6,42
	Nitrat	0,98	4,11	0,53	1,48
	Nitrit	0,0097	0,7	0,0097	0,0097
	Total Coliform	110	4900	1300	3900
	Amonia	0,01	0,3	0,02	0,38
Sei Kera	TSS	1	33	5	21
	pH	6,72	7,46	6,86	7,88
	BOD	1,85	2,99	2,14	2,97
	COD	23,2	28,1	22,4	24,2
	DO	4	7,8	4,84	6,42
	Nitrat	1,27	4,12	0,6	1,48
	Nitrit	0,0097	0,06	0,0097	0,07
	Total Coliform	700	5400	2100	4900
	Amonia	0,04	4,59	0,12	392

Lampiran VI PP Nomor 22 Tahun 2021

1. Kelas satu merupakan air yang peruntukannya dapat digunakan untuk air baku air minum, dan/atau peruntukan lain yang mempersyaratkan mutu air yang sama dengan kegunaan tersebut.
2. Kelas dua merupakan air yang peruntukannya dapat digunakan untuk prasarana/sarana rekreasi air, pembudidayaan ikan air tawar, peternakan, air untuk mengairi pertanaman, dan/atau peruntukan lain yang mempersyaratkan mutu air yang sama dengan kegunaan tersebut.
3. Kelas tiga merupakan air yang peruntukannya dapat digunakan untuk pembudidayaan ikan air tawar, peternakan, air untuk mengairi tanaman, dan/atau peruntukan lain yang mempersyaratkan mutu air yang sama dengan kegunaan tersebut.
4. Kelas empat merupakan air yang peruntukannya dapat digunakan untuk mengairi pertanaman dan/atau peruntukan lain yang mempersyaratkan mutu air yang sama dengan kegunaan tersebut.

Tabel tersebut merangkum hasil pemantauan kualitas air pada tiga sungai, yaitu Sei Batuan, Sei Belawan, dan Sei Kera, yang diukur di segmen hulu dan hilir selama bulan Maret hingga Agustus. Setiap baris memuat identitas lokasi (sungai dan posisi hulu/hilir) serta bulan pengambilan sampel, diikuti parameter TSS, pH, BOD, COD, DO, Nitrat, Nitrit, Total Coliform, dan Amonia, lalu diakhiri dengan kelas mutu air. Secara umum pH berada dalam kisaran sekitar 6,13–8,18, TSS berkisar 1–38 mg/L, BOD sekitar 1,43–2,99 mg/L, COD sekitar 20,8–28,1 mg/L, DO sekitar 4,0–7,8 mg/L, Nitrat kurang lebih 0,48–4,12 mg/L, Nitrit 0,0097–0,7 mg/L, Total Coliform 110–5.400 MPN/100 mL, dan Amonia 0,01–4,69 mg/L dengan satu nilai yang tampak menyimpang sangat tinggi pada salah satu pengukuran di Sei Kera hulu bulan Mei. Berdasarkan kolom kelas, mayoritas sampel tergolong Kelas II, menunjukkan mutu yang relatif baik pada sebagian besar titik pantau. Pada Sei Batuan segmen hulu seluruh bulan berada pada Kelas II, sedangkan di segmen hilir umumnya Kelas II namun turun menjadi Kelas III pada bulan Mei dan Juli, yang berkorelasi dengan lonjakan Total Coliform hingga 5.400 MPN/100 mL dan peningkatan beban amonia. Pada Sei Belawan, baik hulu maupun hilir konsisten berada pada Kelas II dengan variasi parameter yang relatif stabil antar bulan. Pada Sei Kera, segmen hulu tetap Kelas II, sementara segmen hilir menurun ke Kelas III pada bulan Juni, Juli, dan Agustus; penurunan pada Juni dan Juli sejalan dengan nilai Total Coliform mencapai 5.400 MPN/100 mL, sedangkan pada Agustus dipicu oleh kenaikan COD hingga sekitar 28,1 mg/L. Secara keseluruhan, pola ini mengindikasikan kualitas air yang umumnya memenuhi Kelas II, dengan indikasi penurunan mutu terutama pada segmen hilir dan pada bulan-bulan tertentu yang dipengaruhi oleh parameter mikrobiologis serta beban organik, sehingga pengendalian sumber pencemar di bagian hilir menjadi perhatian utama.

3.2 Analisis Data

3.2.1 Pembagian Data Training dan Data Testing

Pada tahap pemodelan, data penelitian dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian dengan perbandingan 80:20. Dengan total 72 observasi, pembagian ini menghasilkan sekitar empat perlima data untuk pelatihan dan satu perlima untuk pengujian; perangkat lunak akan melakukan pembulatan otomatis sehingga jumlah baris tetap 72. Pembagian dilakukan secara acak namun terstratifikasi berdasarkan variabel kelas agar proporsi Kelas II dan Kelas III tetap sebanding pada kedua subset, serta menggunakan penetapan nomor acak yang tetap agar hasil dapat direplikasi. Data pelatihan dipakai untuk membangun dan menyetel model Random Forest di RapidMiner, sedangkan data pengujian yang tidak pernah dilihat model digunakan untuk menilai kinerja di luar sampel. Skema 80:20 atau 58 data training dan 14 data testing dipilih karena ukuran data relatif terbatas tetapi tetap menyediakan cukup data uji untuk mengestimasi performa model secara adil, sekaligus mencegah kebocoran informasi dengan memastikan seluruh proses pra pemrosesan dan penyetelan parameter dilakukan hanya pada data pelatihan sebelum model diterapkan ke data pengujian.

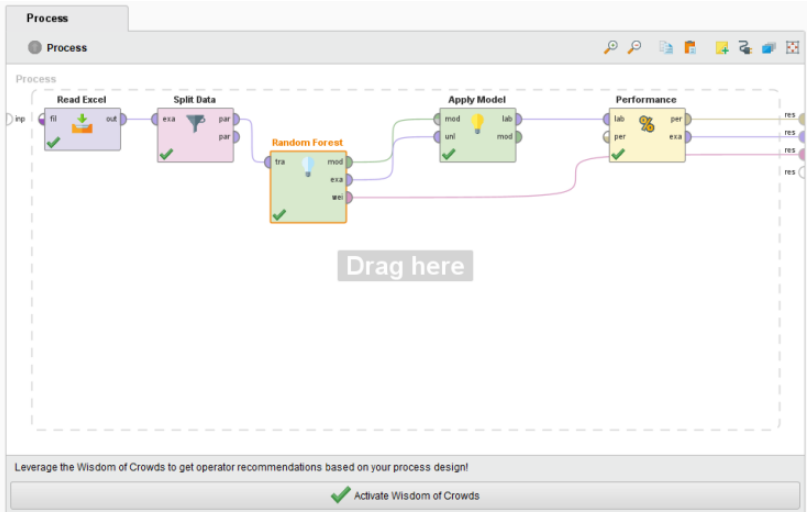
3.2.2 Random Forest Data Training

Proses alur pelatihan model klasifikasi kualitas air di rapidminer dengan skema pembagian data 80:20.

1. Tahap pertama adalah Read Excel yang memuat seluruh dataset kualitas air dari berkas sumber sehingga menghasilkan examples berisi atribut prediktor seperti TSS, ph, BOD, COD, DO, Nitrat, Nitrit, Total Coliform, Amonia serta atribut target Kelas.
2. Output dari pembacaan data kemudian masuk ke Split Data untuk memisahkan data menjadi dua subset, yaitu data training sekitar 80 persen dan data testing sekitar 20 persen. Pada tahap ini disarankan memilih

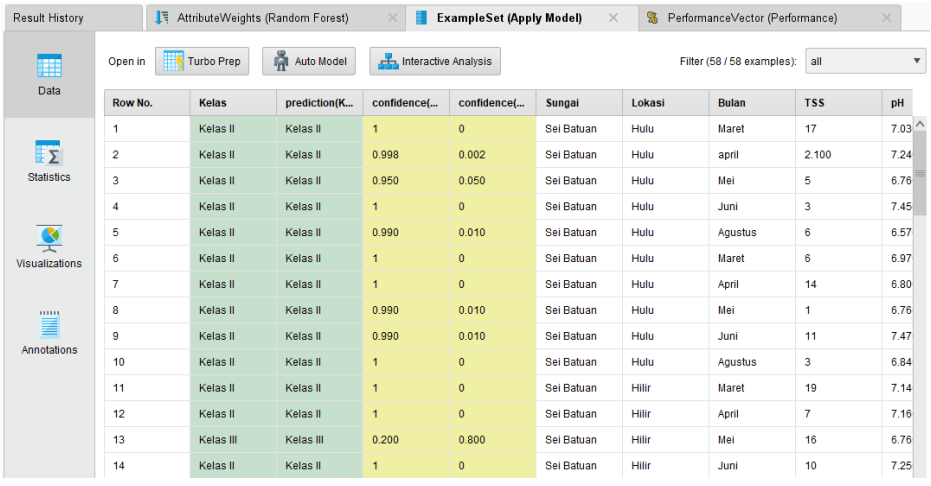
tipe pengambilan sampel stratified dan mengaktifkan penetapan seed acak agar proporsi kelas tetap seimbang di kedua subset dan hasil dapat direplikasi.

3. Subset training dari Split Data dialirkan ke operator Random Forest melalui port training. Di tahap ini algoritma membangun sejumlah 100 pohon keputusan dari sampel acak fitur dan baris pada data training, lalu menggabungkan hasil voting dari seluruh pohon untuk mempelajari pola pemisahan antar kelas mutu air.
4. Hasil pelatihan berupa model terlatih keluar melalui port model, sedangkan aliran data uji dari Split Data masuk ke Apply Model melalui port unlabeled untuk diberi prediksi oleh model tersebut.
5. Keluaran Apply Model adalah data testing yang sudah memiliki kolom prediksi kelas, dan inilah yang dikirimkan ke operator Performance. Pada tahap evaluasi, Performance menghitung ukuran kinerja seperti akurasi, precision, recall, F1, serta menyajikan confusion matrix sehingga kualitas generalisasi model dapat dinilai tanpa bias karena data testing tidak pernah digunakan pada saat pelatihan.

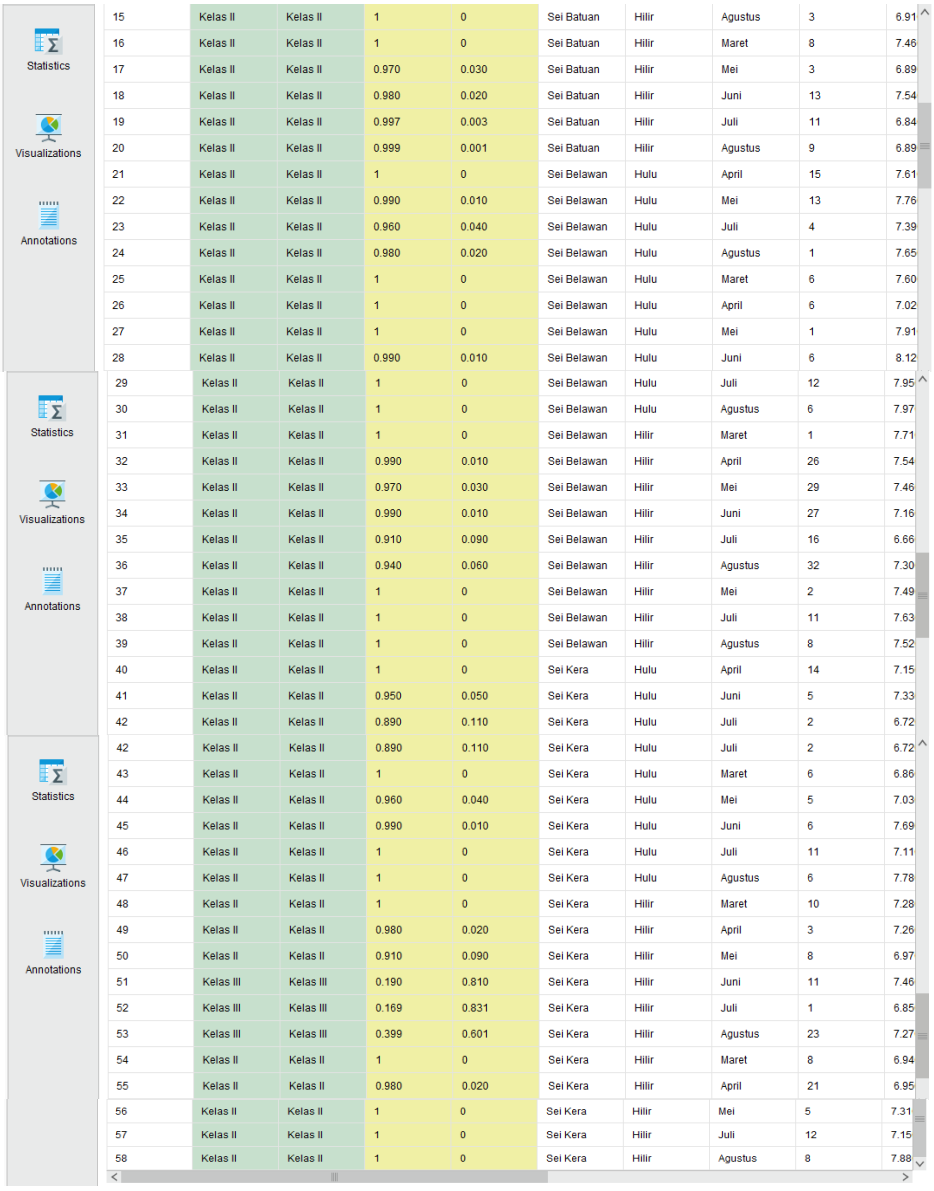


Gambar 1. Alur Pelatihan Model Data Training

Karena fokusnya adalah data training, inti pengerjaan terletak pada aliran dari Split Data ke Random Forest. Subset 80 persen inilah yang dipakai menyesuaikan parameter internal hutan keputusan seperti jumlah pohon, kedalaman maksimum, dan ukuran daun minimum. Praktik ini mencegah kebocoran informasi dan memastikan bahwa pengukuran kinerja di tahap Performance merefleksikan kemampuan model pada data baru. Sebagai penguatan, pastikan atribut Kelas sudah disetel sebagai label sebelum pelatihan, tangani nilai hilang pada atribut numerik maupun kategorikal bila ada, dan gunakan penetapan seed agar hasil pelatihan konsisten ketika proses dijalankan ulang.

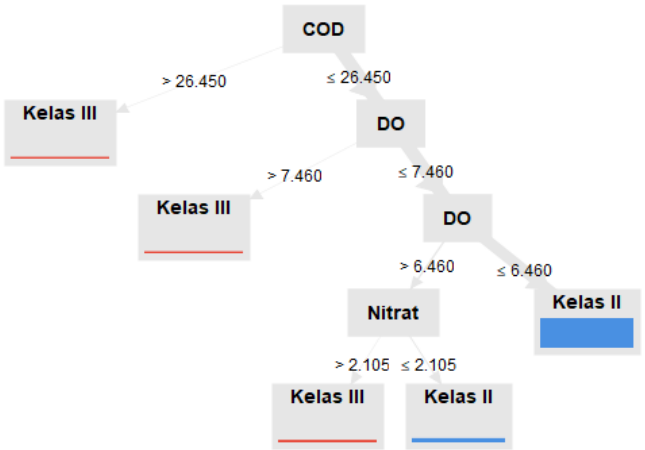


Row No.	Kelas	prediction(K...	confidence(...	confidence(...	Sungai	Lokasi	Bulan	TSS	pH
1	Kelas II	Kelas II	1	0	Sei Batuan	Hulu	Maret	17	7.03
2	Kelas II	Kelas II	0.998	0.002	Sei Batuan	Hulu	april	2.100	7.24
3	Kelas II	Kelas II	0.950	0.050	Sei Batuan	Hulu	Mei	5	6.76
4	Kelas II	Kelas II	1	0	Sei Batuan	Hulu	Juni	3	7.45
5	Kelas II	Kelas II	0.990	0.010	Sei Batuan	Hulu	Agustus	6	6.57
6	Kelas II	Kelas II	1	0	Sei Batuan	Hulu	Maret	6	6.97
7	Kelas II	Kelas II	1	0	Sei Batuan	Hulu	April	14	6.80
8	Kelas II	Kelas II	0.990	0.010	Sei Batuan	Hulu	Mei	1	6.76
9	Kelas II	Kelas II	0.990	0.010	Sei Batuan	Hulu	Juni	11	7.47
10	Kelas II	Kelas II	1	0	Sei Batuan	Hulu	Agustus	3	6.84
11	Kelas II	Kelas II	1	0	Sei Batuan	Hilir	Maret	19	7.14
12	Kelas II	Kelas II	1	0	Sei Batuan	Hilir	April	7	7.16
13	Kelas III	Kelas III	0.200	0.800	Sei Batuan	Hilir	Mei	16	6.76
14	Kelas II	Kelas II	1	0	Sei Batuan	Hilir	Juni	10	7.25

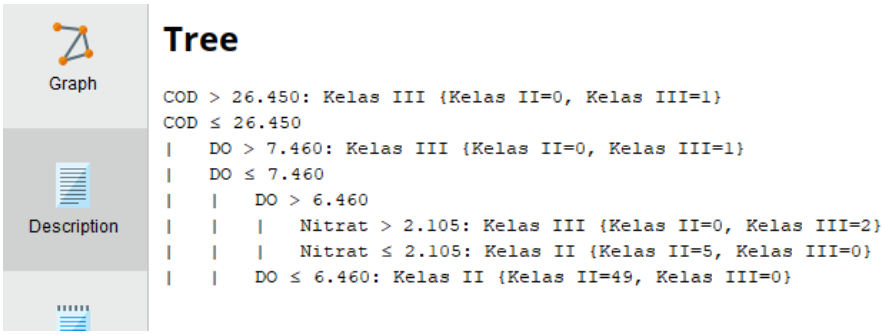


15	Kelas II	Kelas II	1	0	Sei Batuan	Hilir	Agustus	3	6.91
16	Kelas II	Kelas II	1	0	Sei Batuan	Hilir	Maret	8	7.46
17	Kelas II	Kelas II	0.970	0.030	Sei Batuan	Hilir	Mei	3	6.89
18	Kelas II	Kelas II	0.980	0.020	Sei Batuan	Hilir	Juni	13	7.54
19	Kelas II	Kelas II	0.997	0.003	Sei Batuan	Hilir	Juli	11	6.84
20	Kelas II	Kelas II	0.999	0.001	Sei Batuan	Hilir	Agustus	9	6.89
21	Kelas II	Kelas II	1	0	Sei Belawan	Hulu	April	15	7.61
22	Kelas II	Kelas II	0.990	0.010	Sei Belawan	Hulu	Mei	13	7.76
23	Kelas II	Kelas II	0.960	0.040	Sei Belawan	Hulu	Juli	4	7.39
24	Kelas II	Kelas II	0.980	0.020	Sei Belawan	Hulu	Agustus	1	7.65
25	Kelas II	Kelas II	1	0	Sei Belawan	Hulu	Maret	6	7.60
26	Kelas II	Kelas II	1	0	Sei Belawan	Hulu	April	6	7.02
27	Kelas II	Kelas II	1	0	Sei Belawan	Hulu	Mei	1	7.91
28	Kelas II	Kelas II	0.990	0.010	Sei Belawan	Hulu	Juni	6	8.12
29	Kelas II	Kelas II	1	0	Sei Belawan	Hulu	Juli	12	7.95
30	Kelas II	Kelas II	1	0	Sei Belawan	Hulu	Agustus	6	7.97
31	Kelas II	Kelas II	1	0	Sei Belawan	Hilir	Maret	1	7.71
32	Kelas II	Kelas II	0.990	0.010	Sei Belawan	Hilir	April	26	7.54
33	Kelas II	Kelas II	0.970	0.030	Sei Belawan	Hilir	Mei	29	7.46
34	Kelas II	Kelas II	0.990	0.010	Sei Belawan	Hilir	Juni	27	7.16
35	Kelas II	Kelas II	0.910	0.090	Sei Belawan	Hilir	Juli	16	6.66
36	Kelas II	Kelas II	0.940	0.060	Sei Belawan	Hilir	Agustus	32	7.30
37	Kelas II	Kelas II	1	0	Sei Belawan	Hilir	Mei	2	7.49
38	Kelas II	Kelas II	1	0	Sei Belawan	Hilir	Juli	11	7.63
39	Kelas II	Kelas II	1	0	Sei Belawan	Hilir	Agustus	8	7.52
40	Kelas II	Kelas II	1	0	Sei Kera	Hulu	April	14	7.15
41	Kelas II	Kelas II	0.950	0.050	Sei Kera	Hulu	Juni	5	7.33
42	Kelas II	Kelas II	0.890	0.110	Sei Kera	Hulu	Juli	2	6.72
42	Kelas II	Kelas II	0.890	0.110	Sei Kera	Hulu	Juli	2	6.72
43	Kelas II	Kelas II	1	0	Sei Kera	Hulu	Maret	6	6.86
44	Kelas II	Kelas II	0.960	0.040	Sei Kera	Hulu	Mei	5	7.03
45	Kelas II	Kelas II	0.990	0.010	Sei Kera	Hulu	Juni	6	7.69
46	Kelas II	Kelas II	1	0	Sei Kera	Hulu	Juli	11	7.11
47	Kelas II	Kelas II	1	0	Sei Kera	Hulu	Agustus	6	7.78
48	Kelas II	Kelas II	1	0	Sei Kera	Hilir	Maret	10	7.28
49	Kelas II	Kelas II	0.980	0.020	Sei Kera	Hilir	April	3	7.26
50	Kelas II	Kelas II	0.910	0.090	Sei Kera	Hilir	Mei	8	6.97
51	Kelas III	Kelas III	0.190	0.810	Sei Kera	Hilir	Juni	11	7.46
52	Kelas III	Kelas III	0.169	0.831	Sei Kera	Hilir	Juli	1	6.85
53	Kelas III	Kelas III	0.399	0.601	Sei Kera	Hilir	Agustus	23	7.27
54	Kelas II	Kelas II	1	0	Sei Kera	Hilir	Maret	8	6.94
55	Kelas II	Kelas II	0.980	0.020	Sei Kera	Hilir	April	21	6.95
56	Kelas II	Kelas II	1	0	Sei Kera	Hilir	Mei	5	7.31
57	Kelas II	Kelas II	1	0	Sei Kera	Hilir	Juli	12	7.15
58	Kelas II	Kelas II	1	0	Sei Kera	Hilir	Agustus	8	7.88

Gambar 2. Hasil Prediksi Klasifikasi Data Training



Gambar 3. Salah Satu Pohon Yang Dihasilkan



Gambar 4. Deskripsi Pohon

Gambar menampilkan keluaran Example Set dari operator Apply Model setelah model Random Forest dilatih dengan jumlah pohon 100 pada subset data pelatihan. Jumlah baris yang terlihat adalah 58 contoh yang sesuai dengan pembagian 80 persen untuk data training. Kolom Kelas menunjukkan label sebenarnya, sedangkan prediction(Kelas) adalah hasil prediksi model. Dua kolom berwarna kuning adalah confidence untuk masing-masing kelas; nilainya merupakan proporsi suara dari 100 pohon yang membentuk hutan. Sebagai ilustrasi, confidence 0,990 berarti 99 dari 100 pohon memilih kelas tersebut, sedangkan 0,010 berarti hanya satu pohon yang memilihnya.

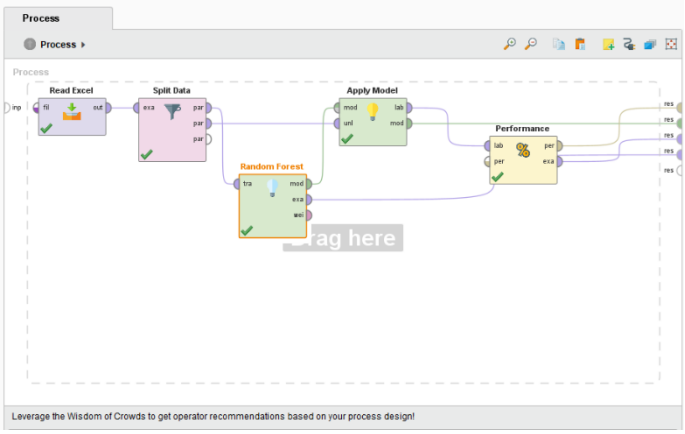
Secara umum, hampir seluruh baris berlabel Kelas II diprediksi sebagai Kelas II dengan confidence sangat tinggi, mayoritas berada pada rentang 0,95 sampai 1,00. Pola ini menunjukkan bahwa kombinasi parameter seperti TSS, BOD, COD, DO, Nitrat, Nitrit, Total Coliform, dan Amonia pada sebagian besar sampel berada cukup jauh dari batas ambang yang memisahkan Kelas II dan Kelas III, sehingga 100 pohon memberikan suara yang konsisten. Pada baris-baris yang berlabel Kelas III, model seringkali juga memberikan prediksi Kelas III dengan tingkat keyakinan yang moderat hingga tinggi, misalnya sekitar 0,60 sampai 0,83, yang menandakan 60 hingga 83 pohon setuju terhadap kelas tersebut.

Penggunaan 100 pohon membuat estimasi probabilitas menjadi cukup halus dengan resolusi 0,01 dan biasanya sudah memadai untuk menstabilkan performa. Banyaknya nilai confidence yang sangat mendekati 1,00 menandakan varians model rendah pada data pelatihan dan pola pemisahan kedua kelas relatif jelas. Walau demikian, karena yang ditampilkan adalah hasil pada data training, kinerja ini cenderung optimistis. Untuk menilai kemampuan generalisasi, penilaian perlu dilihat pada 20 persen data pengujian beserta confusion matrix-nya. Secara ringkas, hasil ini menunjukkan bahwa dengan 100 pohon Random Forest mampu mengenali mayoritas sampel Kelas II dengan keyakinan sangat tinggi dan menangkap sebagian besar sampel Kelas III dengan keyakinan cukup kuat.

3.2.3 Random Forest Data Testing

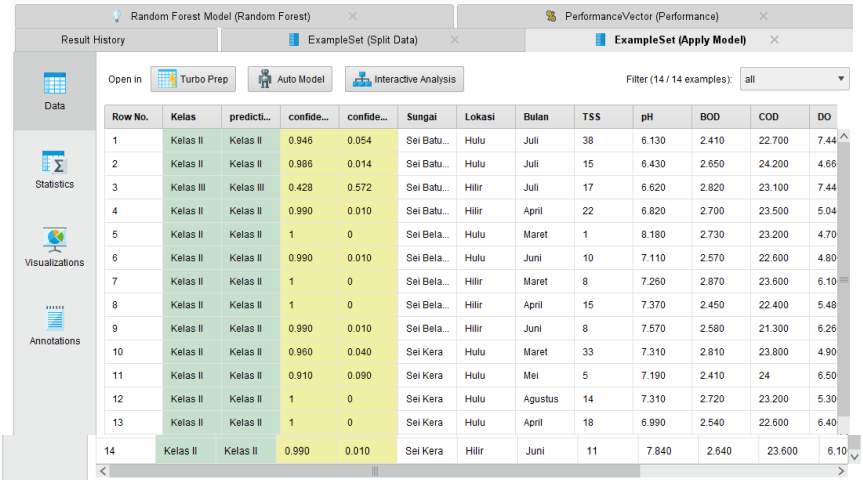
Proses klasifikasi kualitas air sungai dengan metode Random Forest dimulai dari :

1. Pembacaan data dalam format Excel yang berjumlah 72 baris.
2. Data tersebut kemudian dibagi menjadi dua bagian menggunakan metode split data dengan perbandingan 80:20, yaitu 80 persen data digunakan sebagai data latih dan 20 persen digunakan sebagai data uji.
3. Random Forest dilatih dengan data latih untuk membangun model klasifikasi yang mampu mengenali pola dari parameter kualitas air seperti TSS, pH, BOD, COD, dan DO. Model yang sudah terbentuk kemudian diaplikasikan pada data uji untuk menghasilkan prediksi kelas kualitas air sungai.



Gambar 5. Alur Pelatihan Model Data Testing

Random Forest dengan 100 pohon bekerja dengan membangun pohon keputusan secara berulang dari subset data yang berbeda. Setiap pohon memberikan hasil prediksi, kemudian seluruh hasil dari pohon digabungkan untuk menghasilkan keputusan akhir berdasarkan suara terbanyak. Dengan jumlah pohon yang cukup banyak, yaitu 100, model memiliki kemampuan lebih baik dalam mengenali pola, mengurangi kesalahan, dan meningkatkan stabilitas hasil klasifikasi.



Row No.	Kelas	predicti...	confide...	confide...	Sungai	Lokasi	Bulan	TSS	pH	BOD	COD	DO
1	Kelas II	Kelas II	0.946	0.054	Sei Batu...	Hulu	Juli	38	6.130	2.410	22.700	7.44
2	Kelas II	Kelas II	0.986	0.014	Sei Batu...	Hulu	Juli	15	6.430	2.650	24.200	4.66
3	Kelas III	Kelas III	0.428	0.572	Sei Batu...	Hilir	Juli	17	6.620	2.820	23.100	7.44
4	Kelas II	Kelas II	0.990	0.010	Sei Batu...	Hilir	April	22	6.820	2.700	23.500	5.04
5	Kelas II	Kelas II	1	0	Sei Bela...	Hulu	Maret	1	8.180	2.730	23.200	4.70
6	Kelas II	Kelas II	0.990	0.010	Sei Bela...	Hulu	Juni	10	7.110	2.570	22.600	4.80
7	Kelas II	Kelas II	1	0	Sei Bela...	Hilir	Maret	8	7.260	2.870	23.600	6.10
8	Kelas II	Kelas II	1	0	Sei Bela...	Hilir	April	15	7.370	2.450	22.400	5.48
9	Kelas II	Kelas II	0.990	0.010	Sei Bela...	Hilir	Juni	8	7.570	2.580	21.300	6.26
10	Kelas II	Kelas II	0.960	0.040	Sei Kera	Hulu	Maret	33	7.310	2.810	23.800	4.90
11	Kelas II	Kelas II	0.910	0.090	Sei Kera	Hulu	Mei	5	7.190	2.410	24	6.50
12	Kelas II	Kelas II	1	0	Sei Kera	Hulu	Agustus	14	7.310	2.720	23.200	5.30
13	Kelas II	Kelas II	1	0	Sei Kera	Hulu	April	18	6.990	2.540	22.600	6.40
14	Kelas II	Kelas II	0.990	0.010	Sei Kera	Hilir	Juni	11	7.840	2.640	23.600	6.10

Gambar 6. Hasil Prediksi Klasifikasi Data Testing

Hasil prediksi pada data uji menunjukkan performa yang baik, karena sebagian besar kelas yang diprediksi sesuai dengan kelas aslinya. Nilai confidence yang dihasilkan sangat tinggi, berkisar antara 0,94 hingga 1,00. Misalnya, pada data dengan TSS 38 dan pH 6,13, model memprediksi kelas II dengan confidence 0,946, sedangkan pada data dengan TSS 18 dan pH 6,99 model memprediksi kelas II dengan confidence 1,00. Nilai confidence yang tinggi menunjukkan bahwa model sangat yakin terhadap keputusan yang dihasilkan.

Evaluasi kinerja model melalui operator performance memperlihatkan bahwa akurasi yang dicapai sangat baik, ditandai dengan konsistensi antara label aktual dan prediksi. Keberhasilan ini didukung oleh penggunaan 100 pohon yang membuat Random Forest mampu menangkap variasi pada data secara lebih komprehensif. Secara keseluruhan, metode Random Forest dengan 100 pohon terbukti efektif untuk mengklasifikasikan kualitas air sungai berdasarkan parameter TSS, pH, BOD, COD, dan DO. Model ini memberikan prediksi yang akurat dengan tingkat keyakinan tinggi, sehingga dapat digunakan sebagai dasar dalam mendukung pemantauan dan pengelolaan kualitas air sungai.

3.2.4 Kontribusi Setiap Variabel

Hasil analisis menggunakan metode Random Forest dengan 100 pohon tidak hanya menghasilkan prediksi, tetapi juga memberikan informasi mengenai tingkat kepentingan masing-masing atribut dalam proses klasifikasi kualitas air sungai. Tingkat kepentingan ini ditunjukkan oleh nilai weight yang menggambarkan kontribusi setiap variabel terhadap pembentukan model.

Tabel 2. Bobot Kontribusi Variabel

Atribut	Bobot (Weight)
Total Coliform	0,277
COD	0,217
DO	0,11
Nitrat	0,119
TSS	0,075
pH	0,048
Lokasi	0,039
Sungai	0,033
Amonia	0,036
BOD	0,026
Nitrit	0,011
Bulan	0,008

Berdasarkan hasil, atribut dengan kontribusi terbesar adalah Total Coliform dengan bobot 0,277. Hal ini menunjukkan bahwa jumlah total coliform sangat berpengaruh dalam menentukan kelas kualitas air sungai, karena coliform sering digunakan sebagai indikator utama pencemaran mikrobiologis. Selanjutnya, variabel COD memiliki bobot tinggi yaitu 0,217, yang berarti konsentrasi kebutuhan oksigen kimia juga sangat berperan dalam mempengaruhi kualitas air, karena menggambarkan jumlah bahan organik yang terlarut. Atribut DO dengan bobot 0,110 dan Nitrat dengan bobot 0,119 juga memiliki pengaruh signifikan. Keduanya menjadi indikator penting karena oksigen terlarut dan kandungan nitrat berhubungan langsung dengan kondisi ekosistem perairan dan tingkat pencemaran.

Variabel lain seperti TSS dengan bobot 0,075, pH dengan bobot 0,048, serta lokasi, sungai, dan amonia dengan bobot sekitar 0,03 sampai 0,04 memiliki kontribusi sedang. Sementara itu, BOD hanya memiliki bobot 0,026, nitrit dengan bobot 0,011, dan bulan dengan bobot 0,008 memberikan pengaruh yang relatif kecil terhadap klasifikasi. Nilai ini menunjukkan bahwa walaupun parameter tersebut tetap diperhitungkan dalam model, pengaruhnya tidak sebesar variabel utama seperti Total Coliform, COD, DO, dan Nitrat. Secara keseluruhan, hasil pembobotan atribut ini memberikan gambaran bahwa kualitas air sungai paling dipengaruhi oleh parameter mikrobiologis dan kimia utama, khususnya Total Coliform dan COD. Dengan demikian, jika dilakukan pengelolaan kualitas air, kedua parameter tersebut sebaiknya menjadi fokus utama pengendalian, diikuti oleh oksigen terlarut dan kandungan nitrat. Informasi ini memperkuat hasil klasifikasi yang telah dilakukan, serta dapat menjadi acuan dalam menetapkan prioritas pemantauan kualitas air sungai.

3.2.5 Akurasi Keseluruhan

Akurasi keseluruhan adalah ukuran yang menunjukkan seberapa baik model dalam memprediksi kelas pada data uji dibandingkan dengan label sebenarnya. Akurasi dihitung dengan cara membandingkan jumlah prediksi yang benar dengan jumlah seluruh data uji, kemudian dikalikan seratus persen. Dalam konteks penelitian ini, akurasi keseluruhan menggambarkan kemampuan metode Random Forest dalam mengklasifikasikan kualitas air sungai berdasarkan parameter fisik, kimia, dan mikrobiologis. Nilai akurasi yang tinggi menunjukkan bahwa model mampu mengenali pola dengan baik dan memberikan hasil prediksi yang konsisten dengan kondisi aktual. Hal ini berarti model tidak hanya bekerja baik pada data latih, tetapi juga memiliki performa yang kuat pada data uji yang belum pernah dikenali sebelumnya. Dengan demikian, akurasi keseluruhan menjadi indikator penting bahwa metode Random Forest dapat digunakan secara andal untuk mendukung analisis dan pemantauan kualitas air sungai.

Tabel 3. Persentase ketepatan klasifikasi

	Data Training	Data testing
Kelas II	100.00%	100.00%
Kelas III	100.00%	100.00%
Akurasi keseluruhan	100.00%	100.00%

Hasil evaluasi model Random Forest pada penelitian ini menunjukkan tingkat akurasi sebesar seratus persen. Hal ini berarti seluruh data uji berhasil diklasifikasikan dengan benar sesuai dengan kelas aslinya. Berdasarkan confusion matrix, terdapat 54 data yang benar diklasifikasikan sebagai kelas II dan 4 data yang benar diklasifikasikan sebagai kelas III tanpa adanya kesalahan prediksi. Tidak ada data yang tertukar antar kelas sehingga model mampu memisahkan kedua kategori dengan sempurna.

Selain akurasi, nilai precision dan recall untuk masing-masing kelas juga mencapai seratus persen. Pada kelas II, seluruh data yang diprediksi sebagai kelas II benar-benar sesuai dengan kelas aslinya, begitu pula dengan kelas III. Demikian juga pada nilai recall, seluruh data kelas II dan kelas III dapat dikenali secara akurat oleh model tanpa adanya kesalahan pengklasifikasian. Hal ini menunjukkan bahwa model tidak hanya mampu memberikan hasil yang tepat, tetapi juga konsisten dalam mengidentifikasi setiap kelas.

Secara keseluruhan, hasil ini membuktikan bahwa penggunaan metode Random Forest dengan 100 pohon pada dataset yang digunakan mampu menghasilkan performa yang sangat baik. Kemampuan model dalam memberikan prediksi sempurna memperlihatkan bahwa variabel-variabel kualitas air yang digunakan sangat representatif dalam membedakan kategori kelas air sungai. Hasil ini juga menunjukkan bahwa Random Forest dapat dijadikan metode yang andal dalam analisis kualitas air karena mampu memberikan tingkat akurasi dan ketepatan prediksi yang sangat tinggi.

4. KESIMPULAN

Kesimpulan penelitian menunjukkan bahwa metode Random Forest mampu mengklasifikasikan kualitas air sungai dengan sangat baik pada data yang digunakan. Model yang dibangun di RapidMiner menggunakan 100 pohon dan skema pembagian 80 : 20 menghasilkan akurasi 100 % pada data uji, yang ditunjukkan oleh confusion matrix tanpa kesalahan prediksi pada kedua kelas. Nilai weight menunjukkan kontribusi relatif tiap fitur terhadap keputusan model di seluruh pohon. Empat fitur teratas adalah Total Coliform 0,277, COD 0,217, Nitrat 0,119, dan DO 0,110; keempatnya menyumbang total 0,723 atau 72,3% pengaruh keputusan. Dua yang paling dominan adalah Total Coliform 27,7% dan COD 21,7% (gabungan 49,4%), diikuti Nitrat 11,9% dan DO 11,0%. Fitur dengan pengaruh menengah meliputi TSS 0,075, sedangkan pengaruh rendah meliputi pH 0,048, Amonia 0,036, Lokasi 0,039, Sungai 0,033, BOD 0,026, Nitrit 0,011, dan Bulan 0,008. Secara praktis, angka-angka ini menegaskan bahwa parameter mikrobiologis dan kimia utama, khususnya Total Coliform dan COD, paling menentukan pembeda antara Kelas II dan Kelas III, sehingga layak dijadikan fokus pemantauan dan pengendalian kualitas air.

DAFTAR PUSTAKA

Alihar, F. (2018). Penduduk dan akses air bersih di kota semarang. *Jurnal Kependudukan Indonesia*, *13*(1), 67-76.

Fauziah, A. (2025). Optimizing Credit Scoring Performance Using Ensemble Feature Selection with Random Forest. *Jurnal Matematika, Statistika dan Komputasi*, *21*(2), 560-572.

Febriawati, L., Mellaty, R., & Widowati, T. (2021). Analisis aksesibilitas air bersih dalam rangka peningkatan ketahanan keluarga di DKI Jakarta. *Jurnal Lemhannas RI*, *9*(2), 24-39.

Indriyani, A. R., Sudarti, S., & Yushardi, Y. (2024). Analisis limbah pencemaran air sungai di kota dan desa. *OPTIKA: Jurnal Pendidikan Fisika*, *8*(1), 29-35.

Januari, A. D., Rusdayanti, N., Kardian, S., & Shara, S. (2024). Urbanisasi Jakarta dan dampaknya terhadap sosial ekonomi dan lingkungan. *Sustainable Transportation and Urban Mobility*, *1*(1).

Liku, A. L. A., Mulya, W., Sari, I. P., Sipahutar, M. K., & Noeryanto, N. (2022). Mengidentifikasi sumber pencemaran air limbah di tempat kerja. *EUNOIA: Jurnal Pengabdian Masyarakat*, *1*(1), 14-19.

Nugroho, J., Zid, M., & Miarsyah, M. (2020). Potensi sumber air dan kearifan masyarakat dalam menghadapi risiko kekeringan di wilayah karst (Kabupaten Gunung Kidul, Provinsi Yogyakarta). *Jurnal Pengelolaan Lingkungan Berkelanjutan (Journal of Environmental Sustainability Management)*, 438-447.

Pratiwi, D. Y. (2020). Dampak pencemaran logam berat terhadap sumber daya perikanan dan kesehatan manusia. *Jurnal Akuatek*, *1*(1), 59-65.

Rosyidah, M. (2018). Analisis Pencemaran Air Sungai Musi Akibat Aktivitas Industri (Studi Kasus Kecamatan Kertapati Palembang). *Jurnal Redoks*, *3*(1), 21-32.

SP, M. S. U., & Nugroho, H. W. (2023, August). Kajian Algoritma C4. 5 dan K-NN Untuk Memprediksi Penduduk Miskin. In *Prosiding Seminar Nasional Darmajaya* (Vol. 1, pp. 231-241).

Sukristiyono, S., Purwanto, R. H., Suryatmojo, H., & Sumardi, S. (2021). Analisis kuantitas dan kualitas air dalam pengembangan pemanfaatan sumber daya air sungai di kawasan hutan lindung Sungai Wain. *Jurnal Wilayah dan Lingkungan*, *9*(3), 239-255.

Suryani, A. S. (2020). Pembangunan air bersih dan sanitasi saat pandemi Covid-19. *Aspirasi: Jurnal Masalah-Masalah Sosial*, *11*(2), 199-214.

Wijaya, Y. N., & Potalangi, J. G. (2024). Kualitas Air Sungai Di Sulawesi Utara: Systematic Literature Review Ditinjau Dari Parameter Fisika, Kimia Dan Biologi. *TEKNO*, *22*(90), 2381-2389.