Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

KLASIFIKASI KUALITAS AIR SUNGAI DAERAH ISTIMEWA YOGYAKARTA (DIY) MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Andika Ristianto[1], Fadil Indra Sanjaya[2]
[1]andikaristianto66@gmail.com, [2]fadil.Indra@staff.uty.ac.id
[1]Program Studi Sains Data, Universitas Teknologi Yogyakarta
[2]Program Studi Informatika, Universitas Teknologi Yogyakarta

Abstrak

Ketersediaan air bersih yang memadai berpengaruh signifikan terhadap kesehatan masyarakat, pertumbuhan ekonomi, dan keberlanjutan lingkungan. Data Badan Pusat Statistik Daerah Istimewa Yogyakarta tahun 2021 menunjukkan peningkatan jumlah pelanggan air bersih sebesar 8,20%. Data Badan Pusat Statistik Daerah Istimewa Yogyakarta tahun 2022 menunjukkan sungai memiliki peran penting sebagai sumber daya air utama untuk perusahaan air bersih. DLHK sebagai dinas terkait menentukan status mutu air dengan menggunakan metode *STORET* atau metode Indeks Pencemaran (IP), yang tentunya memerlukan waktu yang lama. Oleh karena itu, diperlukan pemantauan yang efisien dalam klasifikasi kualitas air sungai untuk mengatasi permaslahan yang ada. Penelitian ini mengembangkan model klasifikasi dengan *Random Forest* menggunakan metode *cross-validation* yang menunjukkan akurasi rata-rata 100% pada data pelatihan dan 91,43% pada data pengujian. Meskipun ada indikasi *overfitting*, model dengan indeks ke-4 dipilih dan menunjukkan akurasi klasifikasi 97,93% pada data baru. Hasil ini menunjukkan bahwa model mampu mengklasifikasi kualitas air sungai di DIY dengan baik.

Kata kunci: Kualitas Air Sungai, *Machine Learning, Python, Random Forest*.

Abstract

The availability of clean water significantly impacts public health, economic growth, and environmental sustainability. Data from the Central Bureau of Statistics of Yogyakarta in 2021 showed an 8.20% increase in clean water customers. Data from 2022 indicated that rivers play a crucial role as the main water resource for supply companies. The DLHK determines water quality status using the STORET or Pollution Index (IP) Method, which is time-consuming. Therefore, efficient monitoring and classification of river water quality are needed. This study developed a classification model using random forest and cross-validation, showing an average accuracy of 100% on training data and 91.43% on testing data. Despite indications of overfitting, the model with the 4th index was selected and showed a classification accuracy of 97.93% on new data. These results indicate the model can effectively classify river water quality in Yogyakarta.

Keywords*: River Water Quality, Machine Learning, Python, Random Forest.*

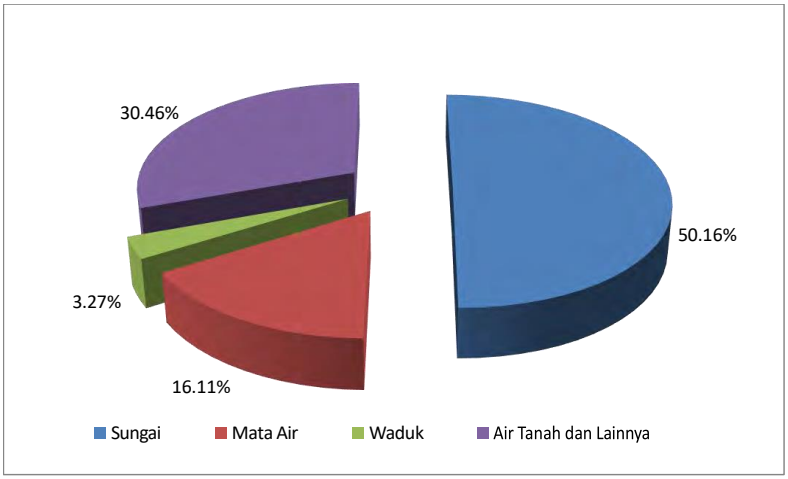
1. Pendahuluan

Sungai merupakan aliran air bersifat permanen atau sementara yang mengalir dari sumber air menuju ke laut, danau, atau sungai lainnya. Peranan air sungai sangat penting dalam mendukung kehidupan organisme air dan manusia, baik sebagai sumber air minum, irigasi pertanian, transportasi, maupun keperluan industri. Air dalam sungai mudah terkena pencemaran oleh limbah dari industri, rumah tangga, pertanian, dan aktivitas manusia lainnya yang dapat menyebabkan penurunan kualitas air sungai, mengancam kelangsungan ekosistem air, serta membahayakan kesehatan manusia yang bergantung pada air tersebut. Masyarakat perlu memiliki pemahaman yang mendalam mengenai kondisi air sungai untuk menjaga keberlanjutan dan melindungi kesehatan. Sungai yang terpapar oleh berbagai bakteri dapat berpengaruh terhadap kualitas sumur milik warga yang berada di sekitaran bantaran sungai. Air bersih

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

yang aman dan berkualitas memiliki peran penting dalam pembangunan dan kesejahteraan manusia serta membangun akses terhadap air bersih merupakan upaya efektif dalam meningkatkan kesehatan dan mengurangi tingkat kemiskinan [1]. Akses air bersih yang baik berpengaruh terhadap pencegahan kontaminasi mikrobiologi penyebab berbagai penyakit diare, kolera, hepatitis A, demam tifoid, dan polio.

Daerah Istimewa Yogyakarta (DIY) yang kaya sumber daya alam terutama air bersih, menghadapi tantangan serius akibat pertumbuhan populasi yang signifikan dan peningkatan aktivitas manusia. Fenomena ini menimbulkan tekanan besar pada kualitas air bersih, suatu aspek krusial dalam kehidupan sehari-hari. Pada data Badan Pusat Statistik Daerah Istimewa Yogyakarta tahun 2021 terjadi peningkatan signifikan jumlah pelanggan air bersih sebesar 8,20% dibandingkan dengan tahun sebelumnya dengan mayoritas pelanggan terdiri dari rumah tangga yang mencapai 197.107 pelanggan pada tahun 2021, dari banyaknya air bersih yang disalurkan atau dikonsumsi juga meningkat dilihat pada tahun 2018 sebesar 32.639 Juta m3 menjadi 37.033 juta m3 pada tahun 2021 yang dapat disimpulkan terjadi pertumbuhan sebesar 3,37% per tahun [2]. Perusahaan-perusahaan penyedia air bersih di Daerah Istimewa Yogyakarta menggunakan air bawah tanah sebagai sumber primer, baik yang terdapat pada lapisan dangkal maupun yang terletak dalam serta memanfaatkan aliran sungai sebagai tambahan. Pada Gambar 1 secara total, persentase terbesar dari sumber-sumber air yang digunakan berasal dari aliran sungai, mencapai 50,16%, diikuti oleh air bawah tanah yang menyumbang sebesar 30,46%, mata air dengan kontribusi sebesar 16,11%, dan waduk dengan jumlah yang relatif kecil, yaitu 3,27% [3].



Gambar 1. Produksi Air Menurut Sumbernya (Sungai, Mata Air, Waduk, Air Tanah dan lainnya) 2022 [3]

Sesuai dengan Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 Dinas Lingkungan Hidup dan Kehutanan (DLHK) Daerah Istimewa Yogyakarta (DIY) menentukan status mutu air dengan menggunakan Metode *STORET* atau Metode Indeks Pencemaran (IP) [4], hal tersebut tentu memerlukan waktu yang lama. Kebutuhan masyarakat terhadap air bersih yang terus meningkat di Daerah Istimewa Yogyakarta (DIY) jika tidak ditangani secara efektif berdampak serius terhadap ekosistem lokal dan meningkatkan beban pada sumber daya air bersih di DIY. Dinas Lingkungan Hidup dan Kehutanan (DLHK) Daerah Istimewa Yogyakarta (DIY) sebagai lembaga terkait memerlukan pendekatan pemantauan yang efektif dan kemampuan klasifikasi kualitas air sungai yang mumpuni. Mengingat sungai adalah sumber air terbesar di daerah tersebut sehingga pemantauan dan klasifikasi yang efektif serta akurat menjadi sangat penting. Dari permasalahan yang dijabarkan, penulis ingin membuat suatu model klasifikasi kualitas air sungai di Daerah Istimewa Yogyakarta (DIY) menggunakan *machine learning* dengan algoritma *Random Forest* untuk memaksimalkan pengelolaan air yang efektif dan akurat.

Dari judul penelitian yang diangkat penulis terdapat beberapa penelitian terdahulu yang melakukan pembuatan model *machine learning* untuk melakukan klasifikasi kualitas air bersih. Salah satunya adalah penelitian yang membahas pentingnya kualitas air tanah sebagai sumber air bersih bagi 32% penduduk DKI Jakarta. Masalah lingkungan dan sanitasi yang buruk menyebabkan penurunan kualitas air tanah, yang sering terkontaminasi oleh bakteri seperti E. coli. Penelitian ini menggunakan algoritma *Naïve Bayes Gaussian* untuk menganalisis data dari 1068 sumber, dan hasilnya divisualisasikan untuk

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

memetakan pencemaran air tanah [5]. Teknik evaluasi melibatkan *cross-validation* dan *percentage split*, dengan akurasi tertinggi sebesar 84,36%. Hasil penelitian ini diharapkan dapat meningkatkan kesadaran masyarakat tentang kualitas air tanah dan membantu perencanaan perlindungan sumber daya air serta menunjukkan bahwa algoritma *Naïve Bayes Gaussian efektif* untuk klasifikasi kualitas air tanah di DKI Jakarta.

Penelitian selanjutnya meneliti pentingnya mengklasifikasi kualitas air bersih untuk memastikan air layak dikonsumsi. Penelitian ini menggunakan metode *Naïve Bayes* untuk memprediksi kualitas air berdasarkan beberapa parameter, seperti rasa, bau, kekeruhan, pH, suhu, sisa klorin, dan *analyzer*. Data yang digunakan meliputi data primer dari wawancara dan observasi, serta data sekunder bersifat privat yang merupakan data perusahaan, terdiri dari 226 data dengan 8 atribut [6]. Hasil penelitian yang dilakukan dengan menggunakan *RapidMiner* menunjukkan akurasi sebesar 97,35%, mengindikasikan bahwa metode *Naïve Bayes* sangat efektif untuk klasifikasi kualitas air bersih. Dengan akurasi yang tinggi, metode ini memiliki potensi besar untuk digunakan dalam evaluasi kualitas air secara efisien, mempermudah dan mempercepat proses penentuan kualitas air dibandingkan dengan metode tradisional yang memerlukan analisis laboratorium yang kompleks.

Berdasarkan penelitian yang membahas tentang pentingnya air bagi kehidupan dan bagaimana kualitas air dapat digunakan sebagai indikator tingkat kesehatan manusia. Penelitian ini fokus pada PDAM Tirta Kencana Kabupaten Jombang dalam rentang waktu tahun 2016 hingga 2017 [7]. Metode yang digunakan untuk menentukan kualitas air bersih menggunakan dua metode yaitu *K-Nearest Neighbor* dan *Naïve Bayes*. Hasil dalam penelitian ini ditemukan bahwa rata-rata nilai akurasi metode *K-Nearest Neighbor* adalah 82,42% sementara metode *Naïve Bayes* memiliki rata-rata nilai akurasi sebesar 70,32%. Dari hasil tersebut, disimpulkan bahwa metode yang paling baik untuk klasifikasi kualitas air bersih pada PDAM Tirta Kencana Kabupaten Jombang adalah metode *K-Nearest Neighbor*.

Penelitian yang mengkaji pentingnya pemeliharaan kualitas air Sungai Citarum, yang sebelumnya terkenal sebagai sungai paling tercemar di dunia. Metode dalam penelitian ini mengaplikasikan teknik *machine learning* untuk mengklasifikasi kualitas air sungai menggunakan metode *K-Nearest Neighbors*, *Support Vector Machine*, dan *Random Forest* [8]. Penelitian ini dilakukan dengan *dataset* dari Dinas Lingkungan Hidup Provinsi Jawa Barat terkait kualitas air Sungai Citarum dari tahun 2018 hingga 2022. Dataset yang digunakan mencakup 8 lokasi sepanjang sungai dengan total 2500 data dan 8 parameter yang berbeda. Hasil penelitian menunjukkan bahwa *Random Forest* memberikan kinerja terbaik dengan akurasi mencapai 99,24%. Ketika dikombinasikan dengan metode *AdaBoost*, akurasi meningkat menjadi 99,34%. Penggunaan teknik *machine learning* terbukti efektif dalam memberikan prediksi yang akurat dan dapat digunakan dalam klasifikasi kualitas air sungai khususnya dalam konteks Sungai Citarum.

Menurut penelitian yang menentukan model melalui proses perbandingan dengan algoritma lain untuk mendapatkan sebuah model prediksi kualitas air sungai yang efektif dan akurat. Penelitian ini membandingan beberapa algoritma dengan menggunakan 8 parameter sebagai variabel prediktor, yaitu *Dissolved Oxygen* (DO), pH, Nitrat, Suhu, Salinitas, *Total Dissolved Solids* (TDS), *Conductivity* (DHL), dan Kekeruhan [9]. Hasil penelitian didapatkan akurasi tertinggi yaitu pada algoritma *JST* (5 *hidden layer*) sebesar 94,6%, diikuti *SVM* 79,3%, *Random Forest* 99,7%, *Naive Bayes* 89,5%. Namun penelitian tersebut hanya menitikberatkan pada evaluasi akurasi model yang didapatkan dari perbandingan algoritma tanpa adanya implementasi langsung dari model yang telah ditentukan sehingga mengakibatkan kurangnya pemanfaatan model oleh masyarakat atau pihak terkait dalam melakukan prediksi kualitas air Sungai Ciliwung.

Penelitian sebelumnya yang bertujuan untuk memprediksi kualitas air sungai di Daerah Istimewa Yogyakarta berdasarkan 7 parameter kualitas air sungai yaitu *Total Solid Suspended* (TSS), *Disolved Oxygen* (DO), *Biochemical Oxygen Demand* (BOD), *Chemical Oxygen Demand* (COD), Total Phosfat (T-P), Fecal Coliform (F.Coli), Total Coliform (T.Coli). Penelitian menggunakan data yang diperoleh dari Dinas Lingkungan Hidup dan Kehutanan (DLHK) Daerah Istimewa Yogyakarta (DIY). Tahapan awal yang dilakukan adalah *data selection*, *preprocessing/cleansing data*, *dan transformation* data untuk memilih serta menyiapkan data sehingga dapat digunakan dalam pembuatan model. Pembentukan model dalam penelitian ini menggunakan *Naïve Bayes Classifier* yang merupakan teknik prediksi berbasis probabilistik sederhana berdasarkan pada penerapan teorema Bayes (Aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat antara variabel prediktor [10]. Hasil dari model *Naïve Bayes* yang dibentuk menggunakan 300 data sampel air sungai menghasilkan akurasi sebesar 71,667%, dengan tingkat presisi dan *recall* sebesar 59,791667% dan 51,665%. Penelitian ini juga

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

mengimplementasikan algoritma *Naive Bayes Classifier* dalam pengklasifikasian pencemaran air sungai menggunakan bahasa pemrograman *PHP* untuk membuat situs web dan *MySQL* sebagai sistem manajemen basis data.

Peneliti menganalisis perbedaan signifikan antara enam penelitian sebelumnya dalam pendekatan metodologi, sumber data, dan hasil yang dicapai. Sebagian besar penelitian menggunakan metode *Naïve Bayes*, sementara hanya dua penelitian menggunakan metode *Random Forest*, meskipun algoritma *Random Forest* menghasilkan akurasi yang tinggi pada kedua penelitian tersebut. Tiga dari enam penelitian yang ditinjau menggunakan data dari Dinas Lingkungan Hidup (DLH), menunjukkan preferensi yang lebih tinggi terhadap sumber data ini. Oleh karena itu, penelitian ini memilih untuk mengeksplorasi data kualitas air sungai dari DLH. Setelah melakukan eksperimen dengan beberapa algoritma, diputuskan bahwa *Random Forest* memiliki akurasi terbaik berdasarkan data yang digunakan. Penelitian ini mengembangkan penelitian sebelumnya dengan menambahkan jumlah parameter menjadi 17 dan sampel data dalam kurun waktu lima tahun terakhir untuk meningkatkan kualitas serta validitas hasil. Penelitian ini memiliki dampak signifikan dalam pengelolaan lingkungan dan sumber daya air, khususnya di Daerah Istimewa Yogyakarta (DIY), serta di daerah lain. Model klasifikasi kualitas air sungai yang akurat dan efektif menggunakan algoritma *random forest* dari penelitian ini dapat memperkuat sistem pemantauan dan pengelolaan sumber daya air sungai dengan lebih efisien. Implikasi praktisnya meliputi kemampuan untuk melakukan intervensi tepat waktu terhadap perubahan kualitas air, meningkatkan kesiapan menghadapi potensi kontaminasi atau pencemaran, serta memberikan dasar yang lebih kuat bagi kebijakan perlindungan lingkungan yang berkelanjutan. Diharapkan penelitian ini dapat memberikan kontribusi positif dalam pengelolaan sumber daya air sungai dan membangun pemahaman yang baik terkait kualitas air sungai di DIY serta diadopsi dalam konteks yang lebih luas untuk meningkatkan pengelolaan air sungai di wilayah lain dengan tantangan serupa.

2. Metode

Pengendalian pencemaran air adalah langkah penting untuk menjaga kualitas air sungai. Standar mutu air sungai ditetapkan untuk membandingkan hasil pemantauan sampel dengan baku mutu yang ditentukan dalam Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003. Terdapat dua metode untuk menentukan status mutu air yaitu dengan metode *STORET* dan Metode Indeks Pencemaran (IP). Kedua metode ini membandingkan data kualitas air dengan kelas standar yang disesuaikan dengan tujuan penggunaannya. metode *STORET* membandingkan data kualitas air dengan standar untuk menentukan status mutu, sementara metode Indeks Pencemaran (IP) menggunakan nilai maksimum dan rata-rata rasio konsentrasi parameter terhadap baku mutunya. Pengklasifikasian mutu air berdasarkan sistem nilai US-EPA membagi mutu air ke dalam empat kelas seperti pada Tabel 1, sedangkan pengelompokkan kelas kualitas air dalam Indeks Pencemaran (IP) dapat dilihat pada Tabel 2.

Tabel 1. Tabel Status Nilai US-EPA Kelas Mutu Air

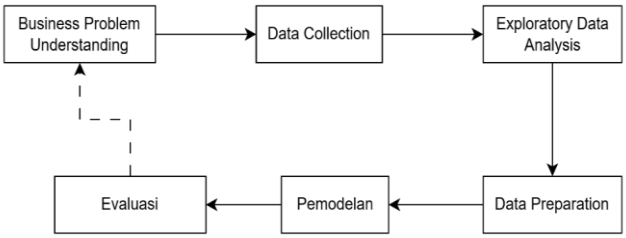
Kelas	Status
Kelas A : Baik Sekali, Skor = 0	Memenuhi Baku Mutu
Kelas B : Baik, Skor = -1 s/d -10	Cemar Ringan
Kelas C : Sedang, Skor = -11 s/d -30	Cemar Sedang
Kelas D : Buruk, Skor > -31	Cemar Berat

Tabel 2. Kategori Status Indeks Pencemaran

Kelas	Status
$0 \leq IP \leq 1{,}0$	Memenuhi Baku Mutu
$1{,}0 < IP \leq 5{,}0$	Cemar Ringan
$5{,}0 < IP \leq 10$	Cemar Sedang
$IP > 10$	Cemar Berat

Tahap penelitian merupakan proses penting yang digunakan penulis dalam menyelesaikan penelitian. Hal ini dilakukan dari awal sampai akhir tahapan agar dapat menyelesaikan penelitian secara sistematis seperti pada Gambar 2.

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575



Gambar 2. Metode Penelitian

2.1. *Business Problem Understanding*

Tahap pertama dalam penelitian ini adalah memahami dan mendefinisikan masalah yang ingin dipecahkan. Kebutuhan bisnis diubah menjadi pertanyaan data sains serta langkah-langkah yang dapat ditindaklanjuti. Cara terbaik melakukannya adalah dengan melibatkan pakar yang memahami masalah tersebut. Berdasarkan pemahaman kebutuhan masyarakat akan air bersih di Daerah Istimewa Yogyakarta (DIY) dan pentingnya sungai sebagai sumber utama air bersih, serta kebutuhan pihak dinas terkait untuk mempercepat proses klasifikasi kualitas air sungai, data hasil pemantauan kualitas air sungai dari Dinas Lingkungan Hidup dan Kehutanan (DLHK) DIY akan digunakan untuk membuat model klasifikasi kualitas air sungai. Model ini bertujuan membantu dalam pemantauan dan pengelolaan sumber daya air sungai secara lebih efisien dan akurat.

2.2. *Data Collection*

Tahap berikutnya adalah pengumpulan data, yang esensial untuk membangun model karena pola-pola yang dibentuk didasarkan pada data yang diperoleh. Data dikumpulkan dari Dinas Lingkungan Hidup dan Kehutanan (DLHK) DIY melalui surat permohonan nomor 1701/F.Saintek-UTY/D/II/2024 dan balasan nomor 000.9/791 yang diteruskan ke Program Pembinaan dan Pengelolaan Lingkungan Hidup (P2KLH). Data yang terkumpul terdiri dari lima *dataset* dalam format pdf berisi hasil pemantauan dan perhitungan Indeks Pencemaran (IP) untuk menentukan kategori kualitas air dari tahun 2019-2023. *Dataset* ini dikonversi ke format *excel* dengan 38 atribut yang diperlukan, meliputi hasil uji dan kategori air berdasarkan IP, dengan jumlah sampel: 2019 sebanyak 151 sampel, 2020 sebanyak 100 sampel, 2021 sebanyak 300 sampel, 2022 sebanyak 150 sampel, dan 2023 sebanyak 150 sampel.

2.3. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) adalah proses memeriksa secara mendalam semua fitur dan properti data, membangun kepercayaan pada data, mendapatkan intuisi, dan mencari cara menangani setiap fitur sebelum membangun model. EDA penting untuk meminimalisir nilai *error* tinggi dan kesalahan prediksi akibat data yang belum siap. Kesalahan seperti *missing value*, *noise*, *inconsistency*, dan duplikasi dapat ditemukan selama EDA. Hasil EDA menunjukkan pola pada *dataset* bertipe numerik dan kategori dengan 38 atribut, termasuk label, dan 851 baris data. Proses EDA dilanjutkan dengan menampilkan distribusi label, *missing values* atribut, dan tren rata-rata IP per tahun untuk analisis data lebih lanjut sebelum pemrosesan. Sampel data keseluruhan yang digunakan dalam penelitian ini dapat dilihat pada Gambar 3.

	Tahun	Bulan	Titik Sample	Suhu	pH	Zat padat terlarut (TDS)	Total suspended solid (TSS)	Oksigen terlarut (DO)	BOD5	COD	...	Tembaga (Cu)	Warna	Debit	Timbal (Pb)	Bakteri Coliform	Bakteri Coli Tinja	Sulfat*	Fenol	IP	Kategori
0	2019	MARET	B-01 Jembatan Sungai Bedog	27.0	7.80	172.8	4.8	7.27	0.10	18.46	...	0.025	32.043	0.36	0.0075	180.0	180.0	NaN	NaN	1.7038	Cemar Ringan
1	2019	MARET	B-02 Jembatan Gamping	29.0	7.14	142.0	6.2	7.27	1.01	6.74	...	0.025	38.000	0.97	0.0075	180.0	180.0	NaN	NaN	1.6520	Cemar Ringan
2	2019	MARET	B-03 Jembatan Kasongan	35.0	6.82	171.0	11.0	7.07	0.61	3.74	...	0.025	47.180	0.63	0.0075	9300.0	1400.0	NaN	NaN	2.2071	Cemar Ringan
3	2019	MARET	B-04 Jembatan Sindon	30.0	6.71	187.0	16.4	7.27	0.81	6.71	...	0.025	58.860	0.706	0.0075	490.0	490.0	NaN	NaN	2.6430	Cemar Ringan
4	2019	MARET	B-05 Jembatan Tempuran Bedog Progo	30.0	7.36	199.0	13.5	8.08	1.82	13.43	...	0.025	81.084	NaN	0.0075	180.0	180.0	NaN	NaN	4.5735	Cemar Ringan
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
846	2023	OKTOBER	TP 4 (KELAS II)	30.0	7.80	107.0	5.0	8.00	2.00	5.00	...	NaN	NaN	NaN	0.0800	240000.0	240000.0	16.0	NaN	9.2245	Cemar Sedang
847	2023	OKTOBER	TP 5 (KELAS II)	29.0	7.40	231.0	3.0	6.00	1.00	3.18	...	NaN	NaN	NaN	0.0800	920000.0	430000.0	19.0	NaN	10.1203	Cemar Berat
848	2023	OKTOBER	TP 6 (KELAS II)	29.0	7.20	233.0	3.0	7.00	1.00	3.18	...	NaN	NaN	NaN	0.0800	920000.0	920000.0	20.0	NaN	11.2895	Cemar Berat
849	2023	OKTOBER	TP 7 (KELAS II)	30.0	7.00	265.0	4.0	5.00	1.00	3.18	...	NaN	NaN	NaN	0.0800	22000.0	9100.0	23.0	NaN	4.1696	Cemar Ringan
850	2023	OKTOBER	TP 8 (KELAS II)	30.0	7.40	286.0	3.0	5.00	1.00	3.18	...	NaN	NaN	NaN	0.0800	43000.0	22000.0	24.0	NaN	5.5304	Cemar Sedang

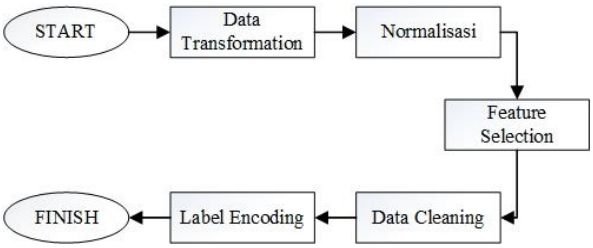
851 rows × 38 columns

Gambar 3. Sampel Data

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

2.4. *Data Preparation*

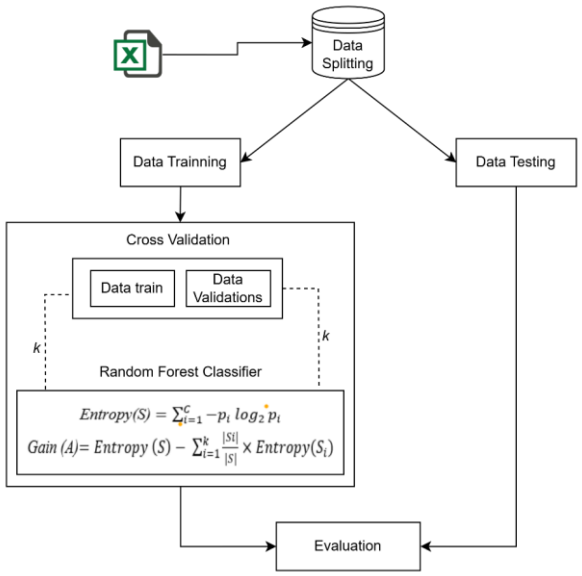
Data preparation adalah proses mengumpulkan, mengolah, dan menyusun data mentah agar siap untuk analisis lebih lanjut, seperti pada Gambar 4. Tahapan ini dimulai dengan memilih data yang relevan serta mengintegrasikan data melalui transformasi dari *dataset* yang berbentuk file pdf menjadi *excel*. Langkah selanjutnya melakukan normalisasi, *feature selection*, *data cleanning*, dan *label encoding* supaya dapat dilakukan analisis lebih lanjut.



Gambar 4. Tahapan *Data preparation*

2.5. *Data Modelling*

Proses dalam langkah ini terdiri dari pemilihan jenis model yang sesuai dan memilih parameter yang terbaik untuk meningkatkan kinerja model dengan tahapan seperti pada Gambar 5. Pemodelan dilakukan dengan membagi data yang dipisahkan menjadi data uji dan data pelatihan, di mana 33% dari data digunakan sebagai data uji untuk menguji kinerja model. Metode perbandingan akurasi kinerja model menggunakan beberapa algoritma klasifikasi digunakan untuk mendapatkan algoritma yang paling cocok dalam klasifikasi data hasil pemantauan kualitas air sungai. Kemudian, model *Random Forest Classifier* dipilih berdasarkan akurasi terbaik yang diperoleh dari proses perbandingan. Penentuan parameter model dilakukan melalui grid parameter untuk mencari parameter yang optimal. *Random Forest Classifier* adalah salah satu algoritma klasifikasi yang termasuk dalam kategori *ensemble learning*. *Ensemble learning* adalah metode yang menggabungkan prediksi dari beberapa model untuk mendapatkan hasil yang lebih baik dibandingkan dengan prediksi dari satu model saja. *Random Forest* menggabungkan beberapa *decision trees* (pohon keputusan) untuk membentuk model yang kuat dan dapat diandalkan. Algoritma ini merupakan kombinasi dari beberapa pohon prediktor, di mana setiap pohon bergantung pada nilai vektor acak yang diambil sampelnya secara independen dan dengan distribusi yang sama untuk semua *tree* [11].



Gambar 5. Tahapan Pemodelan

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

2.6. *Model Evaluation*

Model yang telah dilatih kemudian dilakukan pengujian kinerjanya menggunakan data *validation* untuk mengukur seberapa baik model dapat membuat prediksi pada data yang belum pernah dilihat sebelumnya. Evaluasi dilakukan dengan data uji yang tidak terlihat sebelumnya oleh model atau dikenal sebagai data validasi. Tahapan evaluasi model ini dilakukan dengan 10 kali sesuai dengan nilai k *folds* (lipatan) yang ditentukan pada *cross validation*. Proses evaluasi krusial karena membantu memastikan generalisasi model ke data baru. Berbagai metrik penilaian, termasuk akurasi, presisi, *recall* (sensitivitas), dan *F1-Score* digunakan untuk mengevaluasi performa model. Evaluasi juga melibatkan analisis hasil prediksi model dibandingkan dengan kenyataan atau *ground truth*. Tujuan utama evaluasi model adalah memilih dan mengembangkan model yang ideal untuk memberikan prediksi yang akurat dan berguna dalam aplikasi praktisnya. Evaluasi yang cermat dan teliti penting untuk memastikan kinerja dan kegunaan model yang dihasilkan.

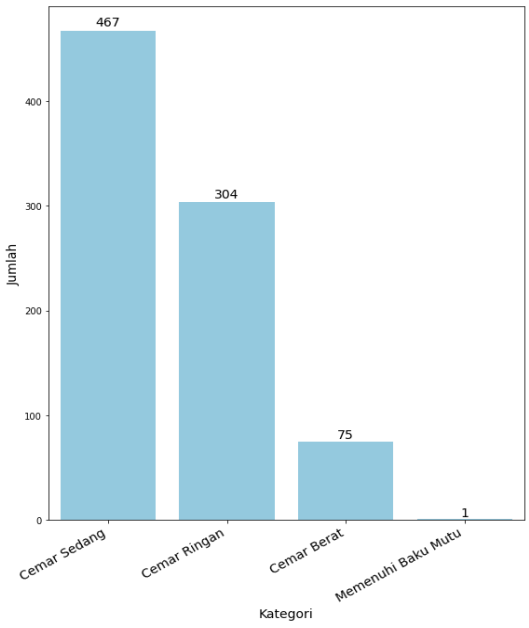
3. Hasil dan Pembahasan

3.1. Hasil

3.1.1. *Exploratory Data Analysis*

1). Distribusi Label

Pada Gambar 6, diagram batang menggambarkan distribusi frekuensi data numerik terkait dengan distribusi label secara visual. Distribusi label menunjukkan bahwa mayoritas sampel dalam *dataset* cenderung memiliki tingkat cemar yang sedang dengan 467 sampel, diikuti oleh tingkat cemar ringan sebanyak 304 label, dan tingkat cemar berat berjumlah 75 sampel. Hasil dalam grafik batang hanya 1 sampel yang memenuhi mutu baku sehingga penulis memutuskan untuk tidak memakai sampel tersebut dalam pembuatan model prediksi. Distribusi ini mengindikasikan bahwa sebagian besar sampel dalam dataset memiliki tingkat cemar yang berkisar antara sedang hingga ringan, sementara tingkat cemar berat dan kualitas yang memenuhi standar baku hanya muncul dalam proporsi yang sangat kecil dari total sampel.



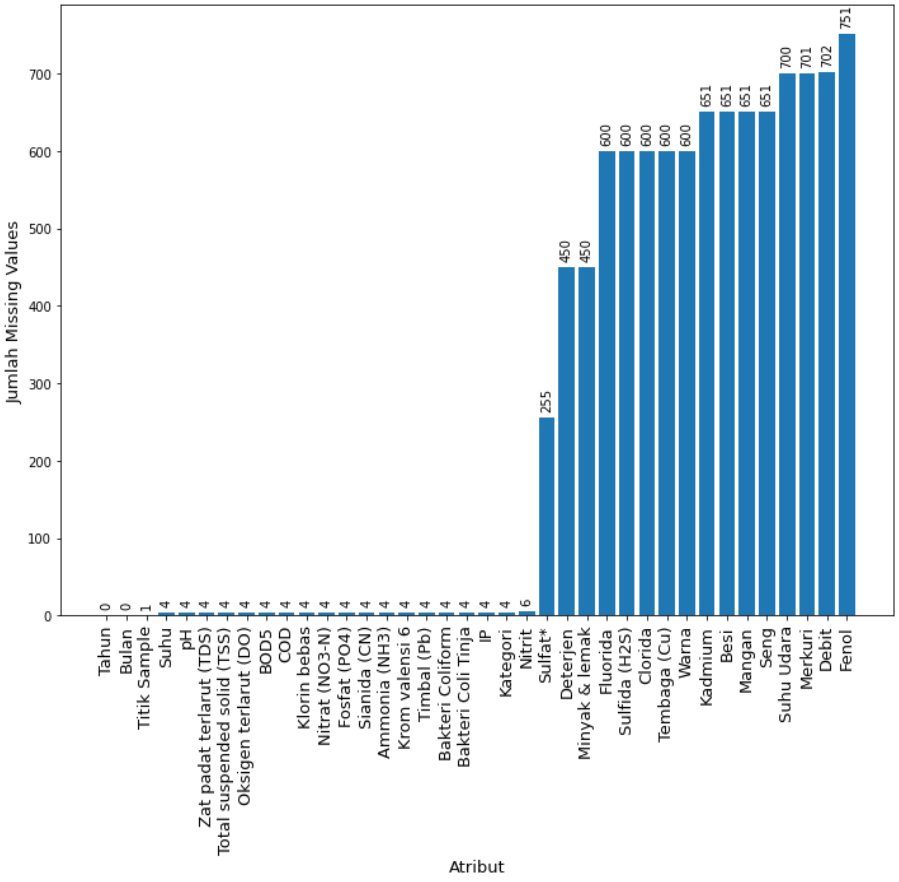
Gambar 6. Grafik Batang Distribusi Label

2). *Missing Values Atribut*

Grafik batang pada Gambar 7, digunakan untuk memberikan gambaran yang jelas tentang sebaran *missing values* pada setiap atribut atau parameter. Bentuk visualisasi sebaran tersebut dapat membantu pemahaman yang lebih baik tentang kondisi data dan memandu proses pengambilan keputusan terkait pemilihan atribut untuk analisis lebih lanjut. Atribut-atribut yang memiliki jumlah *missing values* yang

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

signifikan antara lain adalah atribut fenol, debit, merkuri, suhu udara, seng, mangan, besi, kadmium, warna, tembaga (Cu), klorida, sulfida (H2S), fluorida, minyak & lemak, deterjen, serta sulfat. Sementara untuk beberapa atribut yang lainnya memiliki *missing values* dengan jumlah yang kecil, sehingga atribut masih dapat digunakan dalam proses pembuatan model walaupun harus dilakukan penghapusan *missing values* terlebih dahulu untuk proses analisis selanjutnya.



Gambar 7. Grafik Batang *Missing Values* Atribut

3.1.2. *Data Preparation*

1). Normalisasi

Normalisasi digunakan untuk memastikan konsistensi label dalam kolom kategori pada *dataframe*. Prosesnya dimulai dengan membuat *dictionary* untuk memetakan berbagai varian penulisan label kategori ke bentuk yang baku dan konsisten. Hasil proses tersebut mengubah penulisan label kategori menjadi bentuk yang seragam, sehingga analisis data dapat dilakukan dengan lebih akurat dan konsisten seperti pada Gambar 8. Proses normalisasi label juga melakukan penghapusan terhadap kategori memenuhi mutu baku yang hanya terdapat satu sampel data.

```
Cemar Sedang    467
Cemar Ringan    304
Cemar Berat      75
Name: Kategori, dtype: int64
```

Gambar 8. Hasil Normalisasi Label

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

2). *Feature Selection*

Feature selection dilakukan dengan mempertimbangkan penghapusan fitur-fitur yang memiliki jumlah *missing values* tinggi. Pendekatan ini diambil untuk meningkatkan kinerja model dengan fokus hanya pada fitur-fitur yang memiliki tingkat keinformatifan yang tinggi, sehingga hasil analisis menjadi lebih akurat dan efisien. Fitur yang dihilangkan dalam *dataset* berjumlah 20 dengan 16 atribut memiliki *missing values* tinggi seperti yang sudah disebutkan pada *Exploratory Data Analysis* (EDA) dan 4 atribut tambahan yang tidak digunakan untuk pembuatan model klasifikasi diantaranya Tahun, Bulan, Titik Sampel, Indeks Pencemaran (IP). Sedangkan untuk fitur yang digunakan sebanyak 18 atribut serta atribut tahun yang hanya digunakan sebagai filter pada *splitting data* seperti yang ditampilkan pada Gambar 9.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 850 entries, 0 to 850
Data columns (total 19 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Tahun                        850 non-null    int64
 1   Suhu                         846 non-null    float64
 2   pH                           846 non-null    float64
 3   Zat padat terlarut (TDS)     846 non-null    float64
 4   Total suspended solid (TSS)  846 non-null    float64
 5   Oksigen terlarut (DO)        846 non-null    float64
 6   BOD5                         846 non-null    float64
 7   COD                          846 non-null    float64
 8   Klorin bebas                 846 non-null    float64
 9   Nitrat (NO3-N)               846 non-null    float64
 10  Nitrit                       844 non-null    float64
 11  Fosfat (PO4)                 846 non-null    float64
 12  Sianida (CN)                 846 non-null    float64
 13  Ammonia (NH3)                846 non-null    float64
 14  Krom valensi 6               846 non-null    float64
 15  Timbal (Pb)                  846 non-null    float64
 16  Bakteri Coliform             846 non-null    float64
 17  Bakteri Coli Tinja           846 non-null    float64
 18  Kategori                     846 non-null    object
dtypes: float64(17), int64(1), object(1)
memory usage: 132.8+ KB
```

Gambar 9. Fitur dalam Penelitian

3). *Data Cleanning*

Data cleaning (pembersihan data) adalah proses identifikasi, koreksi, atau penghapusan data yang tidak akurat, tidak lengkap, tidak relevan, atau terduplikasi dalam sebuah *dataset*. Tujuannya adalah untuk memastikan kebersihan dan kualitas data sebelum data tersebut digunakan untuk analisis atau pemodelan. Proses pembersihan data meliputi langkah-langkah seperti mendeteksi dan mengatasi *missing values* seperti pada Gambar 10. Langkah selanjutnya memperbaiki kesalahan penulisan atau format pada nama atribut, serta menghapus 151 data duplikat. Proses pembersihan data yang tepat, dapat meningkatkan validitas dan keandalan hasil analisis yang akan dilakukan.

```
Tahun                        0
Suhu                         4
pH                           4
Zat padat terlarut (TDS)     4
Total suspended solid (TSS)  4
Oksigen terlarut (DO)        4
BOD5                         4
COD                          4
Klorin bebas                 4
Nitrat (NO3-N)               4
Nitrit                       6
Fosfat (PO4)                 4
Sianida (CN)                 4
Ammonia (NH3)                4
Krom valensi 6               4
Timbal (Pb)                  4
Bakteri Coliform             4
Bakteri Coli Tinja           4
Kategori                     4
dtype: int64
```

Gambar 10. *Missing Values*

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

4). *Split Data*

Proses *split* data atau pemisahan data dilakukan untuk membagi *dataset* menjadi dua bagian yang terpisah, yaitu data *train* (latih) dan data *test* (uji) seperti pada tahapan pemodelan. Data latih dibentuk dengan mengambil data dari *dataset* yang sudah dibersihkan dengan filter data kualitas air sungai tahun 2019 hingga tahun 2022 untuk proses pembuatan model menggunakan *cross validation*. Data uji terdiri dari data kualitas air sungai pada tahun 2023 yang nantinya digunakan dalam proses pengujian model. Kolom 'Tahun' dihapus dari kedua *dataset* tersebut karena tidak akan digunakan dalam proses pembelajaran model. Tahap ini bertujuan untuk mempersiapkan data latih dan data uji yang akan digunakan dalam pembuatan serta pengujian model machine learning. Dengan pemisahan ini, model yang dibangun dapat diuji performanya dengan menggunakan data yang belum pernah dilihat sebelumnya.

5). *Label Encoding*

Label encoding yang merupakan proses dalam analisis data untuk mengubah nilai-nilai kategori atau label menjadi representasi numerik. Label encoding digunakan untuk mengubah label kategori menjadi representasi numerik sehingga memungkinkan model *Random Forest Classifier* dapat memproses data dengan efisien. Proses *label encoding* dalam penelitian ini untuk mengubah distribusi label 'Cemar Ringan', 'Cemar Sedang', dan 'Cemar Berat' menjadi angka 0, 1, dan 2, secara berturut-turut. Tujuannya adalah untuk mengubah data kategori menjadi bentuk angka sehingga memungkinkan model untuk memahami dan menganalisis hubungan antara label dan fitur-fitur lainnya. Hasil *label encoding* dapat dilihat pada Gambar 11 dengan rincian jumlah setiap kategorinya.

```
1    295
0    204
2     49
Name: Kategori, dtype: int64
```

Gambar 11. Hasil *Label Encoding*

3.1.3. *Data Modelling*

Tahapan *data modelling* dimulai dengan *cross validation* menggunakan metode *Stratified K-Fold* menggunakan 10 lipatan. Parameter *n_splits* digunakan untuk menentukan jumlah lipatan, sedangkan *random_state* digunakan untuk mengontrol *randomization* dan *shuffle* supaya terjadi pengacakan data sebelum pembagian lipatan. Model *Random Forest Classifier* diinisialisasi dengan parameter *max_depth=15*, yang mengontrol kedalaman maksimum dari pohon keputusan yang akan dibuat. Berdasarkan hasil skor f1 micro yang ditampilkan pada Gambar 12 model mampu menghasilkan prediksi dengan akurasi yang sangat tinggi selama tahap pelatihan. Skor f1 micro menunjukkan nilai konsisten dengan nilai akurasi 100% untuk setiap lipatan yang mengindikasikan bahwa model berhasil belajar dengan baik dari data latih.

```
array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.])
```

Gambar 12. Akurasi *Tranning* Model Setiap Lipatan

Rata-rata akurasi *training* model pada data pelatihan menunjukkan nilai yang sangat tinggi untuk setiap metrik evaluasi seperti pada gambar 13. Nilai rata-rata *untuk F1 Micro Score, Precision Micro Score, Recall Micro Score,* dan *Accuracy Score* adalah 100%, yang menunjukkan bahwa model telah mencapai kinerja yang sempurna pada data pelatihan. Standar deviasi menunjukkan bahwa nilai-nilai evaluasi dari setiap lipatan validasi *cross-validation* memiliki tingkat konsistensi yang tinggi, dengan setiap lipatan memberikan hasil yang sama.

```
Average Train F1 Micro Score: 1.0000 (+/- 0.0000)
Average Train Precision Micro Score: 1.0000 (+/- 0.0000)
Average Train Recall Micro Score: 1.0000 (+/- 0.0000)
Average Train Accuracy Score: 1.0000 (+/- 0.0000)
```

Gambar 13. Rata-Rata Akurasi *Tranning* Model

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

3.1.4. *Model Evaluation*

Hasil akurasi evaluasi model yang dilakukan menggunakan *cross-validation* dengan 10 lipatan menunjukkan variasi nilai akurasi pada setiap lipatan. Berdasarkan Gambar 14 akurasi model bervariasi antara 0.83636364 hingga 0.96363636, dengan sebagian besar nilai akurasi berkisar di sekitar 0.92727273. Nilai-nilai akurasi ini mencerminkan kinerja model yang konsisten dan cukup tinggi dalam melakukan prediksi pada data latih, meskipun terdapat sedikit variasi antara lipatan-lipatan yang berbeda. Evaluasi menggunakan *cross-validation* memberikan gambaran yang lebih komprehensif mengenai kemampuan generalisasi model terhadap data yang tidak terlihat selama pelatihan.

```
array([0.83636364, 0.92727273, 0.87272727, 0.87272727, 0.96363636,
       0.96363636, 0.92727273, 0.92727273, 0.90740741, 0.94444444])
```

Gambar 14. Akurasi Evaluasi Model Setiap Lipatan

Hasil evaluasi pada gambar 15 menunjukkan bahwa model memiliki rata-rata skor F1 Micro, *Precision Micro, Recall Micro,* dan Akurasi sebesar 91,43%, dengan nilai standar deviasi sebesar 3,98%. Model memiliki performa yang konsisten dan stabil dalam melakukan prediksi pada data uji yang tidak digunakan sebelumnya. Skor yang tinggi juga menandakan bahwa model mampu menghasilkan prediksi dengan akurasi yang tinggi secara konsisten.

```
Average F1 Micro Score: 0.9143 (+/- 0.0398)
Average Precision Micro Score: 0.9143 (+/- 0.0398)
Average Recall Micro Score: 0.9143 (+/- 0.0398)
Average Accuracy Score: 0.9143 (+/- 0.0398)
```

Gambar 15. Rata-Rata Akurasi Evaluasi Model

Berdasarkan hasil dari evaluasi model, penulis memutuskan untuk memilih model dengan indeks ke-4. Model ini menunjukkan performa yang sangat baik dengan akurasi 100% pada data train dan 96,36% pada data *test.* Pilihan ini didasarkan pada akurasi kinerja model yang tinggi pada data uji sehingga menunjukkan kemampuan model untuk memprediksi dengan baik pada data yang belum pernah dilihat sebelumnya. Pada Gambar 16 ditampilkan aturan keputusan dari pohon keputusan (*rules tree*) pertama (*tree* 0) dalam model yang terpilih menunjukkan bagaimana berbagai fitur dan ambang batasnya digunakan untuk membuat prediksi. Sebagai contoh, jika nilai *Biochemical Oxygen Demand* (BOD5) kurang dari atau sama dengan 5.05 dan nilai Bakteri Coliform Tinja kurang dari atau sama dengan 11500, pohon melanjutkan ke fitur berikutnya, yaitu oksigen terlarut (DO). Jika nilai DO kurang dari atau sama dengan 9.67 dan Sianida (CN) kurang dari atau sama dengan 0.03, proses dilanjutkan dengan memeriksa Bakteri Coliform dan seterusnya. Setiap cabang pohon berisi kondisi yang lebih spesifik hingga mencapai *node* daun, yang memberikan prediksi berdasarkan nilai-nilai yang diamati. Sebagai contoh, jika nilai Sianida (CN) kurang dari atau sama dengan 0.00, maka pohon ini akan menghasilkan prediksi [[0, 3, 0]]. Aturan-aturan ini digunakan untuk menentukan hasil klasifikasi berdasarkan kondisi yang terpenuhi oleh data input, menciptakan jalur yang berbeda-beda untuk mencapai keputusan akhir.

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

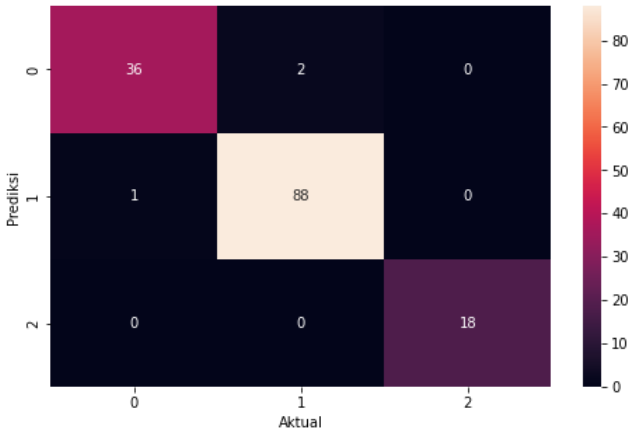
```
Rules for tree 0:
  if BOD5 <= 5.05:
    if Bakteri_Coli_Tinja <= 11500.00:
      if Oksigen_terlarut_(DO) <= 9.67:
        if Sianida_(CN) <= 0.03:
          if Bakteri_Coliform <= 68500.00:
            if Total_suspended_solid_(TSS) <= 28.47:
              if pH <= 7.03:
                if Sianida_(CN) <= 0.00:
                  return [[0. 3. 0.]]
                else:  # if Sianida_(CN) > 0.00
                  if Fosfat_(PO4) <= 0.56:
                    if Zat_padat_terlarut_(TDS) <= 172.00:
                      return [[1. 0. 0.]]
                    else:  # if Zat_padat_terlarut_(TDS) > 172.00
                      return [[0. 2. 0.]]
                  else:  # if Fosfat_(PO4) > 0.56
                    return [[4. 0. 0.]]
              else:  # if pH > 7.03
                if Timbal_(Pb) <= 0.04:
                  if Nitrat_(NO3-N) <= 6.85:
                    if Bakteri_Coli_Tinja <= 4550.00:
                      return [[27.  0.  0.]]
                    else:  # if Bakteri_Coli_Tinja > 4550.00
                      if Bakteri_Coli_Tinja <= 5100.00:
...
          return [[0. 0. 9.]]

================================================================================
```

Gambar 16. Sampel Rules Tree

3.2. Pembahasan

Proses implementasi model dilakukan dengan menguji model menggunakan data kualitas air sungai di Daerah Istimewa Yogyakarta (DIY) tahun 2023. Peneliti memilih model dengan metode *cross-validation* pada indeks ke-4 karena hasil akurasi data pelatihan mencapai 100% dan akurasi data pengujian sebesar 96,36%. Pemilihan model didasarkan pada jarak akurasi yang tidak terlalu jauh antara data pelatihan dan pengujian, dengan tujuan untuk meminimalisir *overfitting*. Setelah model dipilih, peneliti melanjutkan proses dengan melakukan prediksi menggunakan data uji untuk memastikan model dapat memprediksi kualitas air dengan tingkat akurasi yang tinggi dan konsisten.

Prosesnya dengan menciptakan dua *dataframe* baru satu untuk prediksi dan satu untuk kategori aktual dari data uji. Kemudian kedua *dataframe* digabungkan dengan menggabungkan fitur-fitur dari data uji, kategori aktual, dan prediksi yang dihasilkan oleh model. Penggabungan data digunakan untuk membandingkan kategori aktual dengan prediksi yang dilakukan oleh model untuk mengevaluasi kinerja model pada data uji. Hasil pengujian model yang telah dilakukan dilanjutkan dengan mencari nilai akurasi model menggunakan data real. *Confusion matrix* juga dibuat untuk dapat menilai kinerja model prediksi yang telah di buat.



Gambar 17. *Confussion Matrix*

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 10, No. 1, Januari 2025
ISSN: 2715-906X (Online)
10.51717/simkom.v10i1.575

Confusion matrix dari hasil pengujian menampilkan prediksi pada setiap kelas seperti pada Gambar 17. Kelas aktual 0 (Cemar Ringan) model berhasil memprediksi dengan benar sebanyak 36 sampel, namun salah memprediksi sebagai kelas 1 (Cemar Sedang) sebanyak 1 kali. Tidak ada prediksi yang salah untuk kelas 2 (Cemar Berat). Sementara itu, untuk kelas aktual 1 (Cemar Sedang), model membuat dua kesalahan dengan memprediksi sebagai kelas 0 (Cemar Ringan), namun berhasil memprediksi dengan benar sebanyak 88 kali. Dengan menganalisis *confusion matrix* dapat melihat kinerja model dalam memprediksi data baru. Akurasi kinerja model dalam memprediksi data baru cukup tinggi sebesar 97,93% yang menandakan model mampu memprediksi data baru dengan baik.

4. Kesimpulan dan Saran

4.1. Kesimpulan

Berdasarkan hasil dan analisis yang telah dilakukan, peneliti berhasil mengembangkan model klasifikasi kualitas air sungai di Daerah Istimewa Yogyakarta (DIY) dengan *machine learning* menggunakan algoritma *Random Forest*. Model dibuat dengan metode *cross-validation* yang menunjukkan nilai akurasi rata-rata sebesar 100% pada data pelatihan dan 91,43% pada data pengujian. Meskipun ada indikasi *overfitting* pada model, penulis memilih model dengan indeks ke-4 dengan akurasi pelatihan 100% dan pengujian 96,36% untuk diuji dengan data baru. Model mampu melakukan prediksi data baru dengan baik dengan mencapai akurasi prediksi tinggi sebesar 97,93% ketika dibandingkan dengan data real. Hal ini menunjukkan bahwa model yang dibuat oleh peneliti mampu mengklasifikasi kualitas air sungai di DIY dengan sangat baik. Akan tetapi terdapat kendala dalam *dataset* yang digunakan, yaitu penghapusan data dengan label memenuhi mutu baku karena sampel data yang sedikit. Meskipun *dataset* diambil dari rentang waktu 4 tahun, jumlah data tetap terbatas dan distribusi label tidak merata, dengan hanya satu sampel yang memenuhi mutu baku dalam kurun waktu tersebut.

4.2. Saran

Dalam penelitian ini tentu masih terdapat kekurangan sehingga penulis memberikan saran untuk pengembangan penelitian di masa depan sebagai berikut:

1. Mengumpulkan lebih banyak data kualitas air sungai dari periode waktu yang lebih panjang atau dari berbagai sumber lain untuk meningkatkan jumlah sampel dan memperbaiki distribusi label.
2. Menerapkan teknik penyeimbangan data untuk menangani ketidakseimbangan distribusi label, seperti *oversampling* atau *undersampling*, agar model dapat belajar lebih baik dari *dataset* yang ada.
3. Menguji model dengan data kualitas air sungai dari daerah lain di Indonesia untuk menguji generalisasi model dan memastikan bahwa model dapat diaplikasikan secara luas.
4. Mengeksplorasi metode model lain atau teknik regularisasi untuk mengurangi risiko *overfitting* dan meningkatkan performa model secara keseluruhan.
5. Bekerjasama dengan instansi pemerintah atau lembaga penelitian yang memiliki data kualitas air lebih komprehensif untuk mendapatkan *dataset* yang lebih representatif dan valid.

Dengan beberapa saran tersebut, diharapkan penelitian lanjutan dapat menghasilkan model yang lebih efektif dan akurat dalam mengklasifikasi kualitas air sungai di berbagai wilayah.

Daftar Pustaka

[1] WHO, "Water Sanitation and Health," *World Health Organization*, 2024. https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/water-safety-and-quality
[2] Badan Pusat Statistik Provinsi Daerah Istimewa Yogyakarta, *Statistik Air Bersih Daerah Istimewa Yogyakarta 2021*. Badan Pusat Statistik, 2022.
[3] Badan Pusat Statistik Provinsi Daerah Istimewa Yogyakarta, *Statistik Air Bersih Daerah Istimewa Yogyakarta 2022*. Badan Pusat Statistik, 2023.
[4] Republik Indonesia, *Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tentang Pedoman Penentuan Status Mutu Air*. Jakarta: Sekretariat Negara. Jakarta, 2003.

[5] A. Fathiarahma, N. Sulistiyowati, T. Ridwan, and A. Voutama, "Klasifikasi Kualitas dan Prediksi Kondisi Air Tanah di DKI Jakarta Menggunakan Algoritma Naïve Bayes," *J. Intell. Syst. Comput.*, vol. 05, no. 02, 2023, doi: 10.52985/insyst.v5i2.325.

[6] Sutisna and M. N. Yuniar, "Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve Baiyes," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 243–246, 2023, [Online]. Available: https://doi.org/10.55338/saintek.v5i1.1383

[7] M. A. Rahman, N. Hidayat, and A. A. Supianto, "Komparasi Metode Data Mining K-Nearest Neighbor dengan Naïve Bayes untuk Klasifikasi Kualitas Air Bersih: Studi Kasus PDAM Tirta Kencana Kabupaten Jombang," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 12, pp. 6346–6353, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[8] D. P. Sugandi, M. Kallista, and I. P. D. Wibawa, "Klasifikasi Kualitas Air Sungai Citarum Menggunakan Metode K-Nearest Neighbors, Support Vector Machine, dan Random Forest," *e-Proceeding Eng.*, vol. 11, no. 115, p. 1974, 2024, [Online]. Available: https://repository.telkomuniversity.ac.id/

[9] M. Haekal and W. C. Wibowo, "Prediksi Kualitas Air Sungai Menggunakan Metode Pembelajaran Mesin: Studi Kasus Sungai Ciliwung," *J. Teknol. Lingkung.*, vol. 24, no. 2, pp. 273–282, Jul. 2023, doi: 10.55981/jtl.2023.795.

[10] L. J. Phinci, "Klasifikasi Pencemaran Air Sungai di Daerah Naive Bayesa. (Skripsi Sarjana, Universitas Teknologi Digital Indonesiaa)," 2020. [Online]. Available: http://eprints.akakom.ac.id/id/eprint/8806

[11] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, [Online]. Available: https://doi.org/10.1023/A:1010933404324