**E-Tech**
Volume 11 Number 01 2023
ISSN: Print 2541-3600 – Online 2621-7759
DOI: 10.1007/XXXXXX-XX-0000-00

Received Month DD, 20YY; Revised Month DD, 20YY; Accepted Month DD, 20yy

*http://ejournal.unp.ac.id/index.php/e-techr*

# Application of Machine Learning Models in Water Quality Classification in Lake Maninjau: Random Forest as the Optimal Solution

*Abdurrahman Niarman, Reni Kurnia, Iswandi
[1]Universitas Negeri Islam Negeri Mahmud Yunus Batusangkar, [2] Universitas Negeri Padang, [3] Universitas Islam Negeri Mahmud Yunus Batusangkar
e-mail: [1]aabniarman@uinmybatusangkar.ac.id [2]renikurnia@fip.unp.ac.id
[3]iswandi@uinmybatusangkar.ac.id

**Abstract**
This research develops a machine learning model to classify water quality in Lake Maninjau using data from the Ministry of Environment and Forestry's Onlimo application. The dataset includes parameters such as temperature, pH, DO, conductivity, TDS, salinity, turbidity, nitrate and ammonium. Four machine learning algorithms were tested: Logistic Regression, SVM, Gradient Boosting, and Random Forest. As a result, Random Forest shows the best performance with an average accuracy of 87.33% and a standard deviation of 6.97%, and a test accuracy of 90.63%. This model is effective in monitoring and managing water quality, supporting authorities in water resource management decision making. This research also shows how the integration of machine learning and IoT can provide practical solutions in environmental monitoring.

**Keywords**: *Water Quality Classification, Machine Learning, Random Forest, Enviromental Monitoring, Lake Maninjau*



## Pendahuluan

1. Latar Belakang

Kualitas air adalah salah satu indikator krusial yang mencerminkan kondisi lingkungan dan ekosistem di sekitarnya. Untuk memastikan kondisi air yang sesuai dengan standar sehingga dapat dikonsumsi manusia, pertanian dan industri maka diperlukan pemantauan kualitas air secara berkala. (Wolfram et al., 2021). Hal ini juga berlaku dengan kondisi air di Danau Maninjau, sebuah danau yang terletak di Kabupaten Agam, Sumatera Barat, yang memiliki peran vital dalam kehidupan masyarakat sekitar dan keanekaragaman hayati. Aktivitas manusia seperti pertanian, pariwisata, industri dan kegiatan domestik sangat mempengaruhi kualitas air di Danau Maninjau (Sudarso et al., 2021).

Pemantauan kualitas air secara konvensional dilakukan dengan cara mengambil sampel air dan dibawa ke laboratorium untuk kemudian dianalisis. Metode ini dapat memberikan hasil yang akurat tetapi tidak dapat memberikan hasil klasifikasi kualitas air secara cepat. Pemerintah melalui Kementerian Lingkungan Hidup dan Kehutanan (KLHK) mengembangkan sebuah sistem dengan nama *Online Monitoring* (ONLIMO) yang dapat memantau kualitas air secara *real-time* dan dapat diakses oleh seluruh masyarakat (Damayanti et al., 2022). Hal ini menjadi solusi atas kurang efektif dan efisiennya pemantauan kualitas air konvensional meskipun terdapat beberapa persyaratan yang harus dipenuhi untuk dapat diterapkannya sistem ini di sumber air.

Sistem ONLIMO akan memantau kualitas air secara kontinu dan mengirimkan data beberapa parameter kualitas air ke *server*. Data yang dikirimkan secara berkala dan kontinu akan terakumulasi menjadi sebuah kumpulan data yang sangat besar dan kemudian dapat dijadikan sebagai sebuah *dataset* guna pengembangan model *machine learning*. Hal ini dikarenakan *machine learning* adalah teknologi yang bersifat *data driven* sehingga data adalah menjadi kunci utama dari pengembangan model yang baik.

ONLIMO menggunakan perhitungan dalam menentukan klasifikasi kualitas air sesuai dengan standar baku mutu air (Keputusan Menteri Negara Lingkungan Hidup Nomor 115, 2003). Meskipun model seperti ini sangat efektif, tetapi *machine learning* dapat melakukan klasifikasi dengan pendekatan yang berbeda. *Machine learning* dapat melihat pola dari parameter-parameter yang kita gunakan sebagai *dataset*, menganalisisnya dan memberikan hasil klasifikasi. Dari sekian parameter yang dapat diakses melalui halaman *website* ONLIMO, terdapat beberapa parameter yang memiliki kontribusi lebih besar dibandingkan parameter lainnya. Misalnya penelitian yang dilakukan oleh (Baek et al., 2020; S et al., 2024) menemukan bahwa parameter pH memiliki pengaruh besar dalam penentuan hasil model *machine learning.*

Penelitian sebelumnya telah mencoba untuk melakukan klasifikasi kualitas air meggunakan *machine learning* seperti yang dilakukan oleh (Ningsih et al., 2024). Peneliti mengembangkan model yang dapat mengklasifikasikan kualitas air dan mendapatkan hasil algoritma *Random Forest*, *K-Nearest Neighbour* dan *Support Vector Machine* (SVM) dengan *Random Forest* memberikan hasil akurasi prediksi paling baik berdasarkan beberapa parameter input. Model *Random Forest* juga menunjukkan kapabilitas yang baik dalam mengklasifikasikan *dataset* air dengan parameter kompleks berdasarkan penelitian yang dilakukan oleh (Alomani et al., 2022). Algoritma lainnya yang banyak digunakan dalam pengembangan model *machine learning* klasifikasi kualitas air adalah *Logistic Regression*, meskipun dari beberapa percobaan algoritma ini tidak memberikan hasil yang lebih baik tetapi algoritma ini memberikan *baseline* yang penting bagi pengembangan model lainnya (Aish et al., 2023). Selain itu, algoritma *gradient boosting* juga memberikan performa yang baik dan mampu mencapai skor F-1 tertinggi sebesar 0.78 dalam mengidentifikasi air yang aman (Patel et al., 2022).

Beberapa algoritma *machine learning* telah diuji dalam penelitian ini untuk mengklasifikasikan kualitas air, yaitu *Logistic Regression*, SVM, *Gradient Boosting*, dan *Random Forest*. Keempat algoritma ini dipilih berdasarkan kriteria tertentu: *Logistic Regression* dipilih karena kesederhanaannya dan kemampuannya menyediakan baseline untuk perbandingan; SVM dipilih karena kemampuannya menangani dataset yang tidak seimbang dengan baik dan kinerjanya yang kuat dalam berbagai kondisi; *Gradient Boosting* dipilih karena kemampuannya meningkatkan akurasi dengan metode *ensemble* yang mengurangi bias dan *variance*; dan *Random Forest* dipilih karena kemampuannya menangani data yang besar dan kompleks serta mencegah *overfitting* melalui penggunaan banyak pohon keputusan. Keempat algoritma ini telah terbukti efektif dalam banyak studi sebelumnya, seperti yang diungkapkan oleh Islam Khan et al. (2022a), menjadikan mereka pilihan yang tepat untuk penelitian ini.

Penelitian ini bertujuan untuk mengembangkan model *machine learning* menggunakan beberapa algoritma untuk mengklasifikasikan kualitas air di Danau Maninjau dan tentu saja model ini dapat digunakan pada sumber air lainnya. Model ini diharapkan dapat memprediksi tingkat pencemaran air, yang dikategorikan ke dalam beberapa tingkat seperti memenuhi baku mutu, cemar ringan, cemar sedang, atau cemar berat, berdasarkan parameter yang di inputkan. Fokus penelitian ini adalah pada pengembangan model *machine learning* untuk meningkatkan akurasi dan efisiensi dalam pemantauan kualitas air, yang diharapkan dapat memberikan kontribusi signifikan dalam pengelolaan sumber daya air dan kebijakan lingkungan.

## 2. Masalah Penelitian

Penggunaan *machine learning* dalam pemantauan kualitas air menawarkan solusi potensial untuk mengatasi keterbatasan metode konvensional. Tantangan utama dalam penelitian ini adalah apakah *machine learning* mampu mengklasifikasikan kualitas air dengan parameter *input* yang terbatas. Dalam kondisi di mana pemasangan alat pemantauan dengan sensor yang lengkap tidak memungkinkan, penting untuk mengetahui apakah algoritma *machine learning* masih dapat memberikan hasil yang akurat dan andal.

Dengan kemampuan *machine learning* yang mampu mengolah dan menganalisis data secara efisien, diharapkan solusi ini dapat memberikan wawasan yang berguna dan membantu pengambilan keputusan yang lebih baik dalam pengelolaan sumber daya air. Penelitian terbaru menunjukkan bahwa pemanfaatan *machine learning* dalam pemantauan kualitas air dapat meningkatkan kecepatan dan akurasi pengukuran (Hayder et al., 2020). Pendekatan ini juga dapat meningkatkan efisiensi biaya dan waktu dalam implementasi sistem pemantauan kualitas air, sekaligus memastikan bahwa tindakan yang tepat dapat diambil untuk menjaga kualitas air dan melindungi kesehatan ekosistem dan manusia.

3. Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan model *machine learning* yang efektif dalam mengklasifikasikan kualitas air di Danau Maninjau pada khususnya serta sumber-sumber air lainnya. Peneliti akan memanfaatkan empat algoritma untuk menganalisis data kualitas air yang diperoleh dari aplikasi ONLIMO milik KLHK. Dengan menggunakan data yang mencakup berbagai parameter kualitas air seperti suhu, pH, oksigen terlarut (DO), konduktivitas, total padatan terlarut (TDS), salinitas, turbiditas, nitrat, dan amonium, penelitian ini berupaya membangun model yang mampu memberikan prediksi akurat terhadap status mutu air. Klasifikasi ini akan membantu dalam menentukan apakah air tersebut tergolong pada status air yang memenuhi baku mutu, cemar ringan, cemar sedang, atau cemar berat.

Selain mengembangkan model yang akurat, penelitian ini juga bertujuan untuk mengevaluasi kinerja berbagai algoritma *machine learning* seperti *Logistic Regression,* SVM*,* dan *Gradient Boosting*. Dengan melakukan perbandingan ini, penelitian ini berharap dapat menentukan algoritma yang paling sesuai dan efektif untuk klasifikasi kualitas air dan memberikan *insight* kepada pembaca sehingga memberikan pandangan dalam penentuan penelitian selanjutnya. Hasil dari penelitian ini diharapkan tidak hanya dapat meningkatkan pemahaman tentang bagaimana *machine learning* dapat diterapkan dalam pemantauan kualitas air, tetapi juga memberikan alat yang praktis bagi pemerintah dan pihak terkait dalam mengelola dan melindungi sumber daya air secara lebih efisien dan tepat waktu.

4. Manfaat Penelitian

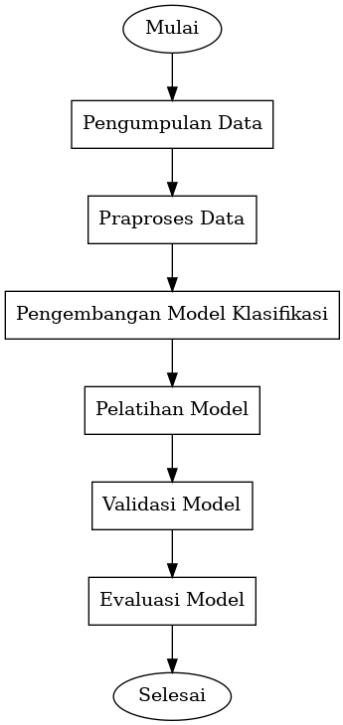
Pengembangan model *machine learning* ini diharapkan dapat memberikan manfaat signifikan dalam pengelolaan dan perlindungan sumber daya air di Danau Maninjau. Dengan menggunakan *machine learning*, model ini dapat berkembang dan meningkatkan akurasinya seiring dengan bertambahnya data dan kompleksitas parameter yang diinputkan. Hal ini berbeda dengan metode konvensional yang bersifat statis dan memerlukan pembaruan manual untuk menangani perubahan dalam data atau kondisi lingkungan. *Machine learning* juga mampu beradaptasi dan belajar dari data baru, sehingga memberikan prediksi yang lebih akurat dan responsif terhadap perubahan kondisi air.

Selain manfaat praktis bagi pengelola sumber daya air, penelitian ini juga memberikan kontribusi akademis dan teknis yang berharga. Dengan mengevaluasi berbagai algoritma *machine learning*, penelitian ini menambah literatur tentang efektivitas dan efisiensi algoritma-algoritma tersebut dalam konteks pemantauan kualitas air. Penemuan ini dapat digunakan sebagai referensi dalam penelitian selanjutnya yang berfokus pada aplikasi teknologi canggih untuk pemantauan lingkungan. Pendekatan yang diambil dalam penelitian ini dapat diadaptasi untuk memantau parameter lingkungan lainnya, seperti kualitas udara di perkotaan, yang dapat membantu dalam upaya konservasi dan pengelolaan sumber daya alam yang lebih luas.

Model *machine learning* ini juga menawarkan keunggulan dalam hal efisiensi biaya dan waktu. Dalam konteks pemantauan kualitas air di daerah dengan sumber daya terbatas, model ini dapat berfungsi dengan baik meskipun hanya menggunakan sejumlah kecil sensor. Hal ini memungkinkan penerapan sistem pemantauan kualitas air yang lebih luas dan merata, termasuk di daerah-daerah terpencil. Dengan kemampuan untuk memproses dan menganalisis data secara cepat dan akurat, model ini memastikan bahwa tindakan yang tepat dapat diambil untuk menjaga kualitas air dan melindungi kesehatan ekosistem dan manusia.

## Metode

Tahapan penelitian ini dilakukan untuk mendapatkan dan mengetahui informasi yang diperlukan dalam proses pengembangan model klasifikasi kualitas air di Danau Maninjau. Informasi yang dibutuhkan dikumpulkan dari berbagai sumber yang relevan seperti jurnal, artikel, buku, dan internet untuk mendukung model yang diusulkan, data yang digunakan, serta tahapan-tahapan yang akan dilakukan. Berikut ini merupakan tahapan penelitian yang meliputi pengumpulan data, *praproses* data, pengembangan dan pelatihan model, validasi model, evaluasi model, hingga penyelesaian akhir.

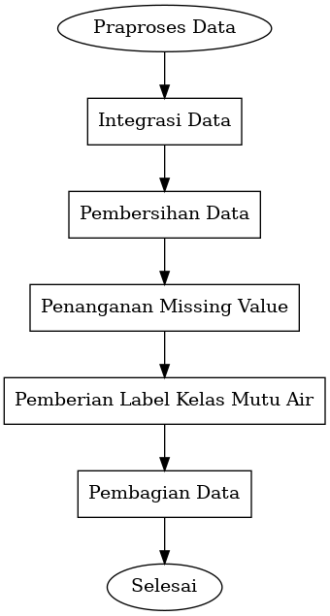



**Gambar 1 *Flowchart* pengembangan model klasifikasi kualitas mutu air**

**1) Pengumpulan Data**

Data yang digunakan dalam penelitian ini diperoleh dari aplikasi ONLIMO milik KLHK, yang dikhususkan pada pengukuran kualitas air di Danau Maninjau. Data yang dikumpulkan terdiri dari berbagai parameter penting kualitas air seperti suhu, pH, oksigen terlarut (DO), konduktivitas, total padatan terlarut (TDS), salinitas, turbiditas, nitrat, dan amonium. Setiap parameter diukur menggunakan sensor yang terpasang di stasiun KLHK Danau Maninjau dan hasilnya dikirimkan ke pusat data setiap jam.

*Praproses* data perlu dilakukan untuk memastikan bahwa data yang digunakan dalam analisis bersih, konsisten, dan bebas dari *error* yang dapat mempengaruhi hasil penelitian. *Praproses* data adalah langkah penting dalam setiap pengembangan model *machine learning* karena kualitas data yang digunakan memiliki dampak langsung pada kinerja model yang dikembangkan (Sami et al., 2021). Data yang tidak bersih dapat menyebabkan kesalahan dalam prediksi dan interpretasi, sehingga proses praproses data menjadi sangat penting untuk memastikan kualitas data yang digunakan, yang pada akhirnya menghasilkan hasil yang valid dan reliabel dalam pengembangan model *machine learning* (Saraswat & Raj, 2022).




**Gambar 2 *Flowchart* praproses data**

Tahapan *praproses* data yang dilakukan meliputi:

1. **Integrasi Data**
   Data yang digunakan dalam penelitian ini berasal dari aplikasi ONLIMO yang dikembangkan oleh KLHK. Data ini mencakup berbagai parameter kualitas air seperti suhu, pH, DO, konduktivitas, TDS, salinitas, turbiditas, nitrat, dan amonium. Integrasi data dalam konteks ini berarti mengumpulkan dan mengorganisir semua data dari ONLIMO menjadi satu *dataset* yang komprehensif dan siap digunakan untuk analisis lebih lanjut.
2. **Pembersihan Data**
   Proses pembersihan data dilakukan untuk menghapus nilai-nilai yang tidak valid atau tidak sesuai, seperti parameter pH yang melebihi batas wajar (> 14). Selain itu, data anomali atau *outlier* yang signifikan diperiksa dan dihapus jika dianggap tidak representatif. Sebagai contoh, data dengan nilai suhu di bawah 0°C atau di atas 100°C, yang tidak mungkin terjadi dalam kondisi alami di Danau Maninjau, dihapus. Langkah ini bertujuan untuk memastikan bahwa hanya data yang valid dan dapat diandalkan yang digunakan dalam analisis lebih lanjut.
3. **Penanganan *Missing Value***
   Mengatasi data yang hilang (*missing values*) dilakukan dengan dua pendekatan utama yaitu imputasi dan penghapusan. Dengan melakukan imputasi, nilai yang hilang diganti dengan nilai *mean* dari parameter terkait. Metode ini digunakan untuk parameter seperti suhu, pH, DO, konduktivitas, TDS, salinitas, turbiditas, nitrat, dan amonium, di mana imputasi *mean* membantu mempertahankan distribusi data asli tanpa memperkenalkan bias signifikan. Kemudian, *record* yang memiliki banyak nilai yang hilang dihapus untuk mencegah analisis yang bias. Sebagai contoh, jika lebih dari 30% data pada satu *record* hilang, *record* tersebut dihapus karena dapat mempengaruhi hasil analisis secara signifikan.
4. **Pemberian Label Kelas Mutu Air**
   Menentukan kelas mutu air berdasarkan metode *Storet* dan baku mutu air kelas 2. Kelas mutu air dikategorikan menjadi empat kategori: memenuhi baku mutu, cemar ringan, cemar sedang, dan cemar berat. Proses ini memberikan label pada setiap data yang digunakan.
5. **Pembagian Data**
   *Dataset* dibagi menjadi *data training* dan *data testing* dengan perbandingan 70% untuk *data training* dan 30% untuk *data testing*. Pembagian ini dilakukan menggunakan metode *stratified sampling* untuk memastikan distribusi kelas yang seimbang di kedua *subset*. Langkah ini penting untuk memastikan bahwa model yang dikembangkan dapat diuji dan divalidasi dengan data yang belum pernah dilihat sebelumnya, sehingga dapat mengukur kinerja model secara objektif.

**2) Pengembangan Model Klasifikasi**

Dalam penelitian ini, empat algoritma machine learning telah dipilih untuk mengembangkan model klasifikasi kualitas air, yaitu *Logistic Regression*, SVM, *Gradient Boosting*, dan *Random Forest*. Pemilihan algoritma ini didasarkan pada beberapa kriteria spesifik yang relevan dengan konteks penelitian ini. Berikut adalah penjelasan dan justifikasi lebih detail mengenai keunggulan spesifik dari setiap algoritma dalam konteks pemantauan kualitas air di Danau Maninjau:

**2.1) *Logistic Regression***

*Logistic Regression* adalah algoritma yang sederhana namun efektif untuk klasifikasi biner dan multi-kelas (Ankrah et al., 2024). Algoritma ini dipilih karena kemampuannya menyediakan *baseline* yang baik dalam mengevaluasi kinerja algoritma lainnya. Keunggulannya terletak pada interpretabilitas dan kemudahan implementasi, menjadikannya pilihan yang solid sebagai model awal untuk klasifikasi kualitas air. Penelitian oleh (Aish et al., 2023) menunjukkan bahwa *Logistic Regression* memberikan wawasan awal yang berguna dalam pengembangan model yang lebih kompleks. Fungsi logit yang digunakan dalam *Logistic Regression* adalah sebagai berikut:

$$\text{logit(p)} = \ln\left(\frac{\text{p}}{1-\text{p}}\right) = \beta_0 + \beta_1 \text{x}_1 + \beta_2 \text{x}_2 + \ldots + \beta_\text{n} \text{x}_\text{n}$$

Di mana $pp$ adalah probabilitas kejadian kelas, $\beta_0\beta_0$ adalah intersep, $\beta_1,\beta_2,\ldots,\beta_n\beta_1,\beta_2,\ldots,\beta_n$ adalah koefisien regresi, dan $x_1,x_2,\ldots,\beta_n x_1,x_2,\ldots,\beta_n$ adalah variabel independen (Wright, 2019).

Dalam penelitian sebelumnya, *Logistic Regression* telah digunakan untuk klasifikasi kualitas air dengan hasil yang bervariasi tergantung pada dataset dan parameter yang digunakan. Sebagai contoh, penelitian oleh (Hassan et al., 2021) menemukan *bahwa Logistic Regression* memiliki akurasi yang cukup baik dalam klasifikasi kualitas air sungai, meskipun tidak selalu menjadi yang terbaik dibandingkan dengan algoritma lain seperti SVM atau *Random Forest*.

**2.2) *Support Vector Machine* (SVM)**

SVM adalah algoritma klasifikasi yang menggunakan *hyperplane* untuk memisahkan data ke dalam kelas yang berbeda antara satu dengan yang lain. Dalam pemantauan kualitas air, SVM berguna untuk mengidentifikasi pola tersembunyi dan menghasilkan prediksi yang tepat. Studi oleh (Suh et al., 2021) menunjukkan bahwa SVM memiliki kinerja unggul dalam *dataset* kualitas air yang tidak seimbang, menegaskan keefektifannya dalam aplikasi ini. Algoritma ini bekerja dengan menemukan *hyperplane* yang memaksimalkan margin antara dua kelas data. Rumus yang digunakan dalam SVM adalah sebagai berikut:

$$w \cdot x - b = 0$$

Di mana $w$ adalah vektor bobot, $x$ adalah vektor fitur, dan $b$ adalah bias. Tujuan dari SVM adalah untuk memaksimalkan margin $\frac{2}{|w|}$ sambil memastikan bahwa data berada di sisi yang benar dari *hyperplane* (Suh et al., 2021).

SVM efektif dalam menangani *dataset* yang kompleks dan tidak seimbang, memaksimalkan margin antar kelas dengan *kernel trick* yang memungkinkan penanganan data kompleks dan *non-linear*. Penelitian oleh (Dogo et al., 2020) menunjukkan bahwa hasil klasifikasi SVM menunjukkan akurasi yang lebih baik dengan tingkat akurasi 92.5% dan *error* klasifikasi 7.5%, dibandingkan dengan KNN yang memiliki akurasi 85.3% dan *error* klasifikasi 14.7%. Ini menunjukkan bahwa SVM lebih unggul dalam memaksimalkan margin klasifikasi dan memberikan hasil yang lebih andal dalam klasifikasi kualitas air..

**2.3) *Gradient Boosting***

*Gradient Boosting* adalah metode *ensemble* yang menggabungkan beberapa model prediktor lemah untuk membentuk model prediktor yang kuat dan efektif dalam meningkatkan akurasi prediksi dengan mengurangi bias dan varians (Zhang et al., 2019). Algoritma ini bekerja dengan membangun model secara bertahap, di mana setiap model baru mencoba mengoreksi kesalahan yang dibuat oleh model sebelumnya. Rumus dasar untuk *Gradient Boosting* adalah sebagai berikut:

$$\widehat{y_i} = \sum_{m=1}^{M} \lambda h_m(x_i)$$

Di mana $\hat{y}_i$ adalah prediksi akhir, $M$ adalah jumlah model, $\lambda$ adalah laju pembelajaran, dan $h_m$ adalah model prediktor lemah ke-m (Zhang et al., 2019).

Dalam konteks klasifikasi kualitas air, *Gradient Boosting* telah menunjukkan performa yang baik. Penelitian oleh (Islam Khan et al., 2022) menggunakan *Gradient Boosting* untuk memprediksi kualitas air sungai dan menemukan bahwa metode ini memberikan hasil yang lebih baik dibandingkan dengan beberapa algoritma lain. *Gradient Boosting* berhasil meningkatkan akurasi prediksi dengan menggabungkan beberapa model sederhana menjadi model yang lebih kompleks dan akurat menjadikannya pilihan yang baik untuk aplikasi ini.

**2.4) *Random Forest***

*Random Forest* adalah algoritma *ensemble* yang menggunakan banyak pohon keputusan untuk klasifikasi. Algoritma ini dipilih karena kemampuannya menangani data besar dan kompleks, serta mencegah *overfitting* (Mohammed & Kora, 2023). Keunggulan *Random Forest* terletak pada ketahanannya terhadap *outliers* dan variabilitas data. Algoritma ini bekerja dengan membangun beberapa pohon keputusan selama pelatihan dan *output* dari kelas yang paling sering muncul (*mode*) dari setiap pohon digunakan sebagai prediksi akhir. Rumus dasar untuk *Random Forest* adalah sebagai berikut:

$$\hat{y} = \text{mode}(\{h_i(x)\}_{i=1}^{n})$$

Di mana $\hat{y}$ prediksi akhir, $h_i(x)$ adalah prediksi dari pohon keputusan ke-i, dan $n$ adalah jumlah pohon keputusan.

*Random Forest* telah terbukti menjadi salah satu algoritma terbaik untuk klasifikasi kualitas air. Penelitian oleh (Victoriano et al., 2020) menunjukkan Model *Random Forest* memiliki performa yang luar biasa dalam klasifikasi kualitas air dengan akurasi mencapai 99.38% dalam studi kasus prediksi polusi sungai di MMORS, Filipina. Penelitian ini menegaskan bahwa *Random Forest* adalah algoritma yang sangat efektif untuk klasifikasi kualitas air, mengingat kemampuannya dalam menangani berbagai jenis data dan memberikan hasil yang akurat dan dapat diandalkan. Penelitian oleh (Alomani et al., 2022) juga menunjukkan bahwa *Random Forest* memberikan akurasi prediksi yang tinggi dalam klasifikasi dataset air dengan parameter yang kompleks.

**3) Evaluasi Model**

Evaluasi model merupakan langkah penting dalam penelitian ini untuk menilai kinerja dan keakuratan model yang dikembangkan. Salah satu metode yang umum digunakan untuk evaluasi model klasifikasi adalah *Confusion Matrix*.

*Confusion Matrix* adalah alat evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dengan membandingkan prediksi model dengan nilai sebenarnya. *Confusion Matrix* terdiri dari empat elemen utama yang mengelompokkan hasil prediksi menjadi empat kategori: *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Berikut adalah tabel yang menggambarkan Confusion Matrix:

**Tabel 1 Tabel *Confusion Matrix***

| ***Actual\Predicted*** | ***Positive (P)*** | ***Positive (P)*** |
|---|---|---|
| *Positive* (P) | *True Positive* (TP) | *False Negative* (FN) |
| *Negative* (N) | *False Positive* (FP) | *True Negative* (TN) |

**Penjelasan Klasifikasi:**

- ***True Positive* (TP)** : Jumlah kasus positif yang diprediksi benar oleh model.
- ***False Positive* (FP)** : Jumlah kasus negatif yang diprediksi salah sebagai positif oleh model.
- ***True Negative* (TN)** : Jumlah kasus negatif yang diprediksi benar oleh model.
- ***False Negative* (FN)** : Jumlah kasus positif yang diprediksi salah sebagai negatif oleh model.

Dari *Confusion Matrix*, kita dapat menghitung beberapa metrik evaluasi untuk menilai kinerja model, yaitu *Accuracy*, *Precision, Recall, Specificity,* dan *F1 Score*.

***Accuracy*** mengukur proporsi prediksi yang benar dari keseluruhan prediksi dan memberikan gambaran umum tentang seberapa baik model dapat memprediksi dengan benar. Dalam konteks klasifikasi kualitas air, *accuracy* penting untuk mengetahui seberapa andal model dalam mengklasifikasikan air ke dalam kategori kualitas yang berbeda secara keseluruhan. *Accuracy* memberikan pandangan umum tentang performa model secara keseluruhan dalam mengklasifikasikan air. Rumus akurasi adalah:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Akurasi merupakan metrik yang paling sederhana dan sering digunakan dalam evaluasi model klasifikasi(Haekal & Wibowo, 2023).

***Precision*** mengukur proporsi prediksi positif yang benar dari keseluruhan prediksi positif. Dalam konteks kualitas air, *precision* penting untuk memastikan bahwa prediksi tersebut akurat ketika model mengidentifikasi air sebagai tercemar. *Precision* membantu mengurangi risiko alarm palsu dalam pengklasifikasian kualitas air.

$$\text{Precision} = \frac{TP}{TP + FP}$$

*Precision* digunakan untuk menilai seberapa tepat prediksi model dalam mengidentifikasi kelas positif (Victoriano et al., 2020).

***Recall*** (*sensitivity*) mengukur proporsi prediksi positif yang benar dari keseluruhan kasus positif sebenarnya. Dalam pemantauan kualitas air, nilai *recall* yang tinggi penting untuk memastikan bahwa model tidak melewatkan banyak kasus air yang benar-benar tercemar. Recall penting untuk memastikan model dapat mendeteksi sebagian besar kasus positif (air tercemar). Rumus *recall* adalah:

$$\text{Recall} = \frac{TP}{TP + FN}$$

*Recall* memberikan informasi tentang kemampuan model dalam menangkap semua kasus positif yang sebenarnya (Victoriano et al., 2020).

***Specificity*** mengukur proporsi prediksi negatif yang benar dari keseluruhan kasus negatif sebenarnya. Nilai *Specificity* yang tinggi memastikan bahwa model tidak memberikan terlalu banyak prediksi positif palsu. Dalam konteks kualitas air, *specificity* memastikan bahwa prediksi negatif (air bersih) adalah benar, menghindari alarm palsu yang menunjukkan air bersih sebagai tercemar. Rumus *specificity* adalah:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

*Specificity* memberikan informasi tentang kemampuan model dalam menghindari kesalahan prediksi positif palsu (Haekal & Wibowo, 2023).

***F1 Score*** adalah rata-rata harmonis dari *precision* dan *recall*. Ini penting dalam konteks kualitas air karena memberikan keseimbangan antara kemampuan model untuk mendeteksi air tercemar (*recall*) dan memastikan bahwa prediksi positif adalah benar (*precision*). *F1 Score* yang tinggi menunjukkan bahwa model tidak hanya baik dalam mendeteksi air tercemar tetapi juga akurat dalam setiap prediksi positifnya, yang penting untuk pengambilan keputusan yang tepat dalam manajemen kualitas air. Rumus *F1 Score* adalah:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

*F1 Score* memberikan gambaran tentang keseimbangan antara kemampuan model untuk memprediksi positif dengan benar dan kemampuan untuk menangkap semua kasus positif yang ada (Victoriano et al., 2020).

## Hasil dan Diskusi

### 1) *Praproses* data

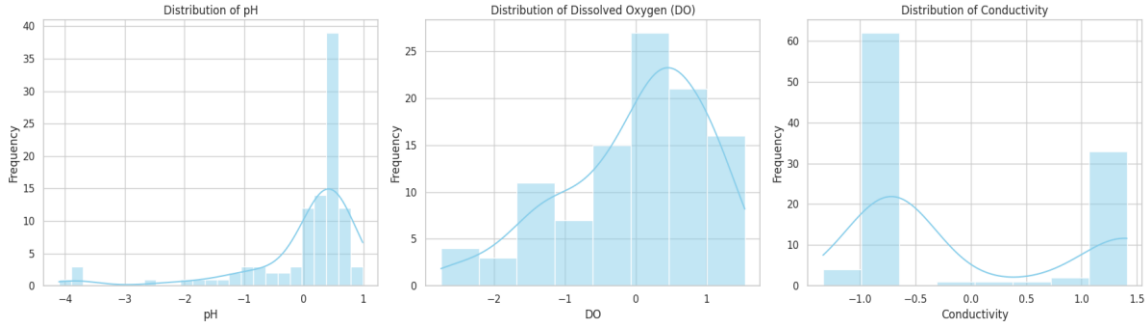
Tahapan *praproses* data dilakukan untuk memastikan data yang digunakan dalam analisis bersih, konsisten, dan bebas dari *error* yang dapat mempengaruhi hasil penelitian. Proses ini meliputi beberapa langkah seperti pembersihan data, imputasi *missing value* dan normalisasi data.

Pembersihan data dilakukan untuk menghapus nilai-nilai yang tidak valid atau tidak sesuai. Berdasarkan analisis awal, ditemukan bahwa parameter pH memiliki nilai yang tidak masuk akal (melebihi 14 atau kurang dari 0). Data yang tidak realistis seperti ini dihapus karena dapat menyebabkan bias dalam model. Selain itu, ditemukan beberapa duplikat dan nilai kosong pada kolom *Nitrate* dan *Ammonium*, yang juga dihapus untuk memastikan integritas *dataset*.

Nilai yang hilang pada kolom *Nitrate* dan *Ammonium* diimputasi menggunakan nilai median karena median tidak terpengaruh oleh *outlier*, sehingga memberikan imputasi yang lebih *robust* dibandingkan dengan *mean*. Selain itu, kolom Status Mutu yang kosong diisi dengan nilai mode ('Cemar Ringan') karena 'Cemar Ringan' adalah kategori yang paling sering muncul.

Setelah melakukan pembersihan dan imputasi pada data, *praproses* data dilanjutkan degan normalisasi data. Normalisasi data dilakukan untuk memastikan bahwa semua parameter memiliki skala yang sama, sehingga tidak ada satu parameter pun yang mendominasi model. Metode normalisasi yang digunakan adalah *Min-Max Scaling*, yang mengubah nilai setiap parameter ke rentang 0 hingga 1. Hal ini penting karena beberapa algoritma *machine learning*, seperti SVM dan *Gradient Boosting*, sensitif terhadap skala data.

Dari delapan parameter yang telah dikoleksi sebagai dataset pada penelitian ini, kami visualisasikan tiga parameter penting yang cukup menarik perhatian kami: pH, *Dissolved Oxygen* (DO), dan *Conductivity*. Kami memilih tiga parameter ini karena ketiga parameter ini memberikan wawasan kritis tentang kualitas air di Danau Maninjau. Histogram dari pH menunjukkan distribusi yang cukup terpusat, yang penting untuk memahami keseimbangan asam-basa air dan dampaknya terhadap ekosistem akuatik. *Dissolved Oxygen* (DO) adalah indikator utama kesehatan air, dengan histogram yang menunjukkan variasi yang signifikan, menyoroti potensi area dengan kondisi *hypoxic* yang dapat mengancam kehidupan akuatik. *Conductivity*, yang memiliki distribusi *bimodal* dalam histogramnya, mengindikasikan adanya dua kondisi berbeda di danau, yang mungkin terkait dengan polusi atau perubahan alamiah dalam lingkungan air.




**Gambar 3 Histogram data pH, DO dan *Conductivity***

Dari grafik histogram di atas dapat kita tarik informasi sebagai berikut :

#### 1.1) pH

Histogram pH menunjukkan bahwa sebagian besar nilai berkonsentrasi di sekitar nilai yang lebih tinggi, namun masih dalam rentang yang dianggap normal untuk air tawar. Hal ini menunjukkan bahwa, meskipun mungkin ada variasi kecil dalam karakter asam atau basa air, air tersebut cenderung stabil dalam hal pH. Hal ini penting dikarenakan pH yang stabil mendukung

kehidupan akuatik yang berkelanjutan dan menunjukkan kurangnya pengaruh polutan asam yang kuat.

**1.2)** ***Dissolved Oxygen***

Distribusi DO memperlihatkan adanya variasi yang cukup lebar dalam konsentrasi oksigen terlarut, yang sangat penting dalam penentuan kualitas air. Air dengan DO rendah bisa menunjukkan kondisi yang tidak mendukung kehidupan ikan dan organisme akuatik lainnya. Penjelasan tentang variasi ini penting karena menunjukkan area atau waktu ketika kualitas air mungkin berada di bawah standar yang ideal.

***1.3) Conductivity***

*Conductivity* menunjukkan dua puncak yang sangat berbeda, mengindikasikan dua kondisi berbeda di dalam danau, yang bisa jadi karena perbedaan sumber air masuk atau perubahan kondisi ekologis. Analisis ini penting karena konduktivitas yang tinggi bisa menunjukkan adanya konsentrasi ion tinggi yang mungkin bersumber dari polusi.

**2) Perbandingan Algoritma yang Dipilih**

Untuk memberikan konteks lebih lanjut, berikut adalah perbandingan hasil dari keempat algoritma yang diuji dalam penelitian ini:

2.1) *Logistic Regression* memberikan akurasi yang cukup baik, dengan *mean accuracy* sebesar 77.62%, *standard deviation* sebesar 10.70%, dan *test accuracy* sebesar 84.38%. Namun, performa ini masih tidak setinggi *Random Forest*. *Logistic Regression* dikenal karena kesederhanaannya dan kemudahan implementasinya, namun cenderung tidak memberikan performa terbaik pada dataset yang kompleks.

2.2) *Support Vector Machine* memberikan hasil klasifikasi yang baik, dengan *mean accuracy* sebesar 80.48%, *standard deviation* sebesar 5.50%, dan *test accuracy* sebesar 84.38%. SVM sangat efektif dalam menangani dataset yang kompleks dan tidak seimbang, serta mampu menemukan *hyperplane* optimal yang memisahkan kelas-kelas data. Namun, SVM memerlukan waktu komputasi yang tinggi untuk dataset besar, yang menjadi salah satu kekurangannya.

2.3) *Gradient Boosting* menunjukkan performa yang kompetitif dengan *mean accuracy* sebesar 79.24%, *standard deviation* sebesar 3.96%, dan *test accuracy* sebesar 84.38%. Algoritma ini efektif dalam meningkatkan akurasi prediksi dengan mengurangi *bias* dan *variance*, menjadikannya pilihan yang baik dalam banyak aplikasi *machine learning*. Namun, Gradient Boosting masih kalah dalam hal performa dibandingkan dengan *Random Forest* dalam penelitian ini

2.4) *Random Forest* menunjukkan hasil yang paling unggul dalam penelitian ini, dengan *mean accuracy* sebesar 87.33%, *standard deviation* sebesar 6.97%, dan *test accuracy* mencapai 90.63%. Selain itu, model ini menunjukkan *precision* sebesar 95.45%, *recall* sebesar 91.30%, *specificity* sebesar 88.89%, dan *F1 Score* sebesar 93.33%. Dengan performa yang sangat baik ini, Random Forest membuktikan dirinya sebagai algoritma yang paling efektif untuk klasifikasi kualitas air di Danau Maninjau. Algoritma ini tidak hanya mampu menangani kompleksitas data, tetapi juga memberikan prediksi yang sangat akurat, menjadikannya pilihan yang optimal untuk aplikasi ini.

**3) Hasil pengembangan Model**

Model klasifikasi kualitas air yang dikembangkan dalam penelitian ini menggunakan algoritma *Random Forest*. Algoritma ini dipilih karena memberikan performa terbaik dibandingkan dengan algoritma lain yang diuji, yaitu *Logistic Regression, Support Vector Machine*, dan *Gradient Boosting*. Penggunaan *Random Fores*t memungkinkan untuk menangani data yang kompleks dan beragam, serta mencegah *overfitting* dengan menggunakan banyak pohon keputusan.

*Random Forest* bekerja dengan membangun sejumlah besar pohon keputusan selama pelatihan dan menggabungkan hasilnya untuk membuat prediksi akhir. Keunggulan utama dari *Random Forest* adalah kemampuannya untuk menangani variabilitas data dan mengurangi risiko *overfitting*, yang sering menjadi masalah pada model *machine learning* yang lebih sederhana. Dalam konteks penelitian ini, *Random Forest* mampu memprediksi kualitas air dengan akurasi yang tinggi, menunjukkan bahwa algoritma ini sangat efektif untuk klasifikasi data yang kompleks seperti beberapa parameter kualitas air yang digunakan pada penelitian ini.

Selain itu, *Random Forest* juga memiliki kemampuan untuk mengukur pentingnya setiap fitur dalam proses klasifikasi. Ini memungkinkan peneliti untuk memahami faktor-faktor mana yang paling berpengaruh terhadap kualitas air di Danau Maninjau. Dengan informasi ini, strategi pengelolaan dan intervensi yang lebih efektif dapat dirancang untuk menjaga kualitas air.

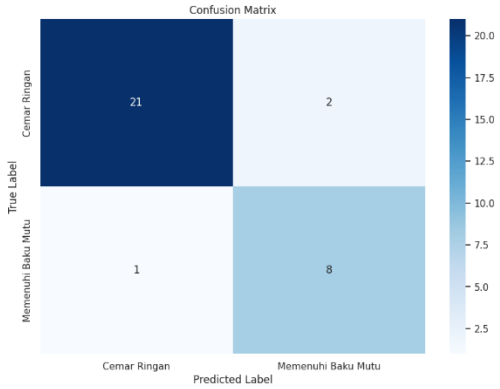
**4) Evaluasi Model dengan *Confusion Matrix***

Evaluasi model dilakukan dengan menggunakan *Confusion Matrix*, yang memberikan gambaran tentang bagaimana model memprediksi nilai-nilai sebenarnya. Berikut kami tampilkan hasil *Confusion Matrix* untuk pengujian model *Random Forest*:

**Tabel 2 Hasil *Confusion Matrix Random Forest***

| ***Actual\Predicted*** | ***Positive* (P)** | ***Negative* (N)** |
|---|---|---|
| ***Positive* (P)** | 21 (TP) | 2 (FN) |
| ***Negative* (N)** | 1 (FP) | 8 (TN) |

*Confusion Matrix* ini menunjukkan bahwa model *Random Forest* memiliki performa yang baik dalam memprediksi kelas positif dan negatif.




**Gambar 4 Hasil *Confusion Matrix Random Forest***

Berikut adalah penjelasan dari setiap elemen dalam *Confusion Matrix* dan hubungannya dengan hasil evaluasi model:

- ***True Positive*** (TP = 21): Ini adalah jumlah kasus di mana model memprediksi positif (misalnya, kualitas air memenuhi baku mutu) dan hasil sebenarnya juga positif. Ini menunjukkan bahwa model *Random Forest* berhasil mengidentifikasi 21 sampel dengan benar sebagai positif.
- ***False Negative*** (FN = 2): Ini adalah jumlah kasus di mana model memprediksi negatif (misalnya, kualitas air tidak memenuhi baku mutu) padahal hasil sebenarnya positif. Ini menunjukkan bahwa ada 2 sampel yang seharusnya positif tetapi diprediksi negatif oleh model.
- ***False Positive*** (FP = 1): Ini adalah jumlah kasus di mana model memprediksi positif padahal hasil sebenarnya negatif. Ini menunjukkan bahwa ada 1 sampel yang seharusnya negatif tetapi diprediksi positif oleh model.
- ***True Negative*** (TN = 8): Ini adalah jumlah kasus di mana model memprediksi negatif dan hasil sebenarnya juga negatif. Ini menunjukkan bahwa model berhasil mengidentifikasi 8 sampel dengan benar sebagai negatif.

*Confusion Matrix* ini penting karena memberikan gambaran detail tentang bagaimana model berperforma dalam hal prediksi yang benar dan kesalahan prediksi. Secara keseluruhan, hasil *Confusion Matrix* menunjukkan bahwa model *Random Forest* memiliki performa yang baik, dengan jumlah *True Positives* dan *True Negatives* yang cukup tinggi serta jumlah *False Positives* dan *False Negatives* yang rendah.

Hal ini mengindikasikan bahwa model *Random Forest* mampu mengklasifikasikan data dengan akurasi yang tinggi dan kesalahan yang minimal, menunjukkan performa yang baik.

**5) Analisis Metrik Evaluasi**

*Accuracy* mengukur proporsi prediksi yang benar dari keseluruhan prediksi. Akurasi memberikan gambaran umum tentang seberapa baik model dapat memprediksi dengan benar. Dalam penelitian ini, akurasi dihitung sebagai berikut:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{21 + 8}{21 + 8 + 2 + 1} = \frac{29}{32} = 0.90625$$

Angka akurasi sebesar 0.90625 menunjukkan bahwa model *Random Forest* mampu memprediksi dengan benar sebanyak 90.63% dari keseluruhan sampel yang diuji. Ini adalah indikator performa yang sangat baik dan menunjukkan bahwa model memiliki tingkat kesalahan yang rendah dalam klasifikasinya.

*Precision* mengukur proporsi prediksi positif yang benar dari keseluruhan prediksi positif. *Precision* penting ketika kesalahan positif dapat berakibat signifikan, misalnya dalam konteks kualitas air yang dapat mempengaruhi keputusan pengelolaan air. *Precision* dihitung sebagai berikut:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{21}{21 + 1} = \frac{21}{22} = 0.9545$$

*Precision* sebesar 0.9545 menunjukkan bahwa dari semua prediksi positif yang dibuat oleh model, 95.45% di antaranya adalah benar positif. Hal ini mengindikasikan model sangat tepat dalam prediksi positifny. Tingginya *precision* mengindikasikan bahwa model jarang memberikan alarm palsu atau *false positives* dalam konteks kualitas air.

*Recall* (*Sensitivity*) mengukur proporsi kasus positif yang benar-benar diprediksi positif oleh model. *Recall* penting dalam situasi di mana mendeteksi semua kasus positif adalah kritis, misalnya dalam memastikan kualitas air tetap baik. *Recall* dihitung sebagai berikut:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{21}{21 + 2} = \frac{21}{23} = 0.9130$$

*Recall* sebesar 0.9130 mengindikasikan bahwa model mampu menangkap 91.30% dari semua kasus positif yang sebenarnya. Ini menunjukkan kemampuan model dalam mendeteksi kasus positif. Tingginya *recall* berarti model jarang gagal mendeteksi kasus positif dalam kualitas air.

*Specificity* mengukur proporsi kasus negatif yang benar-benar diprediksi negatif oleh model. *Specificity* penting dalam konteks mengidentifikasi kasus negatif. *Specificity* dihitung sebagai berikut:

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{8}{8 + 1} = \frac{8}{9} = 0.8889$$

*Specificity* sebesar 0.8889 berarti model mampu dengan benar mengidentifikasi 88.89% dari semua kasus negatif yang sebenarnya. *Specificity* penting dalam menghindari *false positives*, yang dalam konteks kualitas air bisa berarti menghindari tindakan yang tidak perlu. Tingginya *specificity* menunjukkan bahwa model jarang membuat kesalahan dalam mengidentifikasi negatif, sehingga keputusan pengelolaan air dapat dilakukan berdasarkan informasi tersebut.

*F1 Score* adalah metrik yang menggabungkan *Precision* dan *Recall* dalam satu nilai harmonis. *F1 Score* digunakan ketika kita ingin mencapai keseimbangan antara *Precision* dan *Recall*. *F1 Score* dihitung sebagai berikut:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \cdot \frac{0.9545 \cdot 0.9130}{0.9545 + 0.9130} = 0.9333$$

*F1 Score* sebesar 0.9333 memberikan keseimbangan antara *precision* dan *recall*, mencerminkan bahwa model *Random Forest* tidak hanya tepat dalam prediksinya tetapi juga mampu menangkap sebagian besar kasus positif yang sebenarnya. Skor F1 ini menunjukkan performa model yang sangat baik, menggabungkan kekuatan kedua metrik tersebut dalam satu nilai yang harmonis. Tingginya F1 Score berarti model memiliki keseimbangan yang baik antara kemampuan mendeteksi kasus positif dan menghindari alarm palsu dalam konteks pengelolaan kualitas air.

**6) Pembahasan dan Diskusi**

Dari keempat algoritma yang dicoba pada penelitian ini, *Random Forest* mampu memberikan performa terbaik. Algoritma ini mampu menangani data yang kompleks dan beragam, seperti data parameter yang menentukan kualitas air di Danau Maninjau pada khususnya dan sumber-sumber air lainnya. Kemampuan *Random Forest* dalam mencegah *overfitting* dan memberikan hasil prediksi yang akurat membuatnya menjadi pilihan yang optimal untuk klasifikasi kualitas air.

Namun, meskipun *Random Forest* unggul dalam pengujian ini, ada beberapa kelemahan dan tantangan yang perlu diperhatikan. Salah satunya adalah kebutuhan sumber daya komputasi yang tinggi untuk melatih model. Proses pelatihan dengan banyak pohon keputusan membutuhkan waktu dan kapasitas komputasi yang signifikan, terutama untuk *dataset* yang besar jika penelitian ini akan diperluas ke *dataset* yang lebih besar dan kompleks untuk menyempurnakan modelnya. Salah satu solusi yang dapat diusulkan adalah penggunaan teknik *parallel processing* atau *distributed computing* untuk mempercepat proses pelatihan model *Random Forest*. Selain itu, optimisasi *hyperparameter* juga dapat dilakukan untuk mengurangi beban komputasi sambil tetap mempertahankan performa model yang tinggi.

Hasil penelitian ini menunjukkan bahwa model *Random Forest* dapat digunakan untuk membantu *monitoring* kualitas air di Danau Maninjau sehingga pengelolaan kualitas air akan menjadi lebih efektif. Dengan akurasi yang tinggi, model ini dapat memberikan informasi yang tepat dan andal kepada pihak berwenang untuk pengambilan keputusan terkait pengelolaan sumber daya air. Keuntungan lain dari penggunaan model *machine learning* seperti *Random Forest* adalah kemampuannya untuk terus belajar dan beradaptasi seiring bertambahnya data, yang memungkinkan model untuk tetap relevan dan akurat dalam jangka panjang.

Untuk penelitian selanjutnya, eksplorasi lebih lanjut penggunaan algoritma lain atau pengintegrasian data *real-time* dari *Internet of Things* (IoT) untuk meningkatkan akurasi model dengan mengoleksi data yang lebih besar sebagai *dataset*. Selain itu, penelitian dapat difokuskan pada optimasi model *Random Forest* dengan teknik-teknik terbaru dalam *machine learning* untuk meningkatkan performa dan efisiensi komputasi. Misalnya, penggunaan model *ensemble* yang lebih kompleks atau integrasi dengan algoritma *deep learning* dapat dipertimbangkan untuk meningkatkan ketepatan dan keandalan prediksi kualitas air.

## Kesimpulan

Penelitian ini bertujuan untuk mengembangkan model klasifikasi kualitas air di Danau Maninjau menggunakan algoritma *machine learning*. Beberapa parameter yang digunakan pada penelitian ini yang kemudian dijadikan sebagai *dataset* adalah suhu, pH, DO, *conductivity*, TDS, *salinity*, *turbidity*, *nitrate*, dan *ammonium* yang didapat dari arsip data aplikasi ONLIMO. Setelah *praproses* data selesai dilakukan, data-set tersebut diuji dengan empat algoritma, yaitu *Logistic Regression, Support Vector Machine* (SVM), *Gradient Boosting*, dan *Random Forest*, dengan hasil algoritma *Random Forest* menunjukkan performa terbaik. Hasil evaluasi menunjukkan bahwa *Random Forest* memiliki akurasi sebesar 90.63%, *precision* sebesar 95.45%, *recall* sebesar 91.30%, *specificity* sebesar 88.89%, dan *F1 Score* sebesar 93.33%. *Confusion Matrix* menunjukkan bahwa model ini mampu memprediksi 21 sampel dengan benar sebagai positif dan 8 sampel sebagai negatif, dengan hanya 1 prediksi positif yang salah dan 2 prediksi negatif yang salah. Hal ini mengindikasikan bahwa *Random Forest* memiliki performa yang sangat baik dalam mengklasifikasikan data kualitas air.

Secara keseluruhan, penelitian ini menunjukkan bahwa model *Random Forest* adalah alat yang sangat efektif untuk klasifikasi kualitas air di Danau Maninjau. Dengan akurasi yang tinggi dan kemampuan untuk menangani data yang kompleks, model ini dapat membantu dalam pengelolaan kualitas air yang lebih baik. Hasil penelitian ini memiliki implikasi praktis yang signifikan, karena model ini dapat digunakan untuk membantu *monitoring* kualitas air di Danau Maninjau sehingga pengelolaan kualitas air akan menjadi lebih efektif. Implementasi model ini dalam sistem *monitoring* kualitas air dapat membantu mendeteksi perubahan kualitas air secara cepat sehingga pihak berwenang dapat secara tepat memberikan tanggapan untuk menjaga

kualitas air. Penelitian ini juga membuka peluang untuk mengembangkan model prediksi yang lebih canggih dengan memanfaatkan data tambahan seperti parameter lingkungan lainnya, data meteorologi, dan data penggunaan lahan di sekitar Danau Maninjau. Integrasi data dari berbagai sumber ini dapat memberikan gambaran yang lebih komprehensif tentang faktor-faktor yang mempengaruhi kualitas air dan membantu dalam merancang intervensi yang lebih efektif.

## Daftar Pustaka

Aish, A. M., Zaqoot, H. A., Sethar, W. A. & Aish, D. A. (2023). Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques. *Water Practice & Technology*, *18*(3), 501–521. https://doi.org/10.2166/wpt.2023.028

Alomani, S. M., Alhawiti, N. I. & Alhakamy, A. (2022). Prediction of Quality of Water According to a Random Forest Classifier. *International Journal of Advanced Computer Science and Applications*, *13*(6). https://doi.org/10.14569/IJACSA.2022.01306105

Ankrah, B., Brew, L. & Acquah, J. (2024). Multi-Class Classification of Genetic Mutation Using Machine Learning Models. *Computational Journal of Mathematical and Statistical Sciences*, *3*(2), 280–315. https://doi.org/10.21608/cjmss.2024.267064.1040

Baek, S.-S., Pyo, J. & Chun, J. A. (2020). Prediction of Water Level and Water Quality Using a CNN-LSTM Combined Deep Learning Approach. *Water*, *12*(12), 3399. https://doi.org/10.3390/w12123399

Damayanti, A. A., Wahjono, H. D. & Santoso, A. D. (2022). Pemantauan Kualitas Air Secara Online dan Analisis Status Mutu Air di Danau Toba, Sumatera Utara. *Jurnal Sumberdaya Alam Dan Lingkungan*, *9*(3), 113–120. https://doi.org/10.21776/ub.jsal.2022.009.03.4

Dogo, E. M., Nwulu, N. I., Twala, B. & Aigbavboa, C. O. (2020). Empirical Comparison of Approaches for Mitigating Effects of Class Imbalances in Water Quality Anomaly Detection. *IEEE Access*, *8*, 218015–218036. https://doi.org/10.1109/ACCESS.2020.3038658

Haekal, M. & Wibowo, W. C. (2023). Prediksi Kualitas Air Sungai Menggunakan Metode Pembelajaran Mesin: Studi Kasus Sungai Ciliwung. *Jurnal Teknologi Lingkungan*, *24*(2), 273–282. https://doi.org/10.55981/jtl.2023.795

Hassan, Md. M., Hassan, Md. M., Akter, L., Rahman, Md. M., Zaman, S., Hasib, K. Md., Jahan, N., Smrity, R. N., Farhana, J., Raihan, M. & Mollick, S. (2021). Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms. *Human-Centric Intelligent Systems*, *1*(3–4), 86. https://doi.org/10.2991/hcis.k.211203.001

Hayder, G., Kurniawan, I. & Mustafa, H. M. (2020). Implementation of Machine Learning Methods for Monitoring and Predicting Water Quality Parameters. *Biointerface Research in Applied Chemistry*, *11*(2), 9285–9295. https://doi.org/10.33263/BRIAC112.92859295

Islam Khan, Md. S., Islam, N., Uddin, J., Islam, S. & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 4773–4781. https://doi.org/10.1016/j.jksuci.2021.06.003

Keputusan Menteri Negara Lingkungan Hidup Nomor 115. (2003). *Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003*. https://dokumen.tips/documents/kepmen-no-115-tahun-2003.html?page=1

Mohammed, A. & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774. https://doi.org/10.1016/j.jksuci.2023.01.014

Ningsih, L., Jaman, J. H., Salam, N. I. & Haikal, M. (2024). Perbandingan Kinerja Algoritma Klasifikasi Status Mutu Air. *Indonesian Journal of Multidisciplinary on Social and Technology*, *2*(1), 72–76. https://doi.org/10.31004/ijmst.v2i1.298

Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S. & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. *Computational Intelligence and Neuroscience*, *2022*, 1–15. https://doi.org/10.1155/2022/9283293

S, S., Tamatgar, N., Dilli, R. & M, K. (2024). Deployment of Random Forest Algorithm for prediction of ammonia in river water. *Proceedings of the 2024 13th International Conference on Software and Computer Applications*, 18–23. https://doi.org/10.1145/3651781.3651811

Sami, O., Elsheikh, Y. & Almasalha, F. (2021). The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120695

Saraswat, P. & Raj, S. (2022). DATA PRE-PROCESSING TECHNIQUES IN DATA MINING: A REVIEW. *International Journal of Innovative Research in Computer Science & Technology*, 122–125. https://doi.org/10.55524/ijircst.2022.10.1.22

Sudarso, J., Tri Suryono, T. S., P. Yoga, G., Imroatusshoolikhah, I., Ibrahim, A., Laela Sari, L. S., Muhammad Badjoeri, M. B. & Octavianto Samir, O. S. (2021). Effect of Anthropogenic Activity on Benthic Macroinvertebrate Functional Feeding Groups in Small Streams of West Sumatra, Indonesia. *Sains Malaysiana*, *51*(11), 3551–3566. https://doi.org/10.17576/jsm-2022-5111-04

Suh, Y. S., Shin, S. K., Baang, D., Seo, S. M. & Lee, J. B. (2021). *A Brief Review of Non-linear Support Vector Machine for Machine Learning Programming*. https://www.kns.org/files/pre_paper/46/21A-011-%EC%84%9C%EC%9A%A9%EC%84%9D.pdf

Victoriano, J. M., Lacatan, L. L. & Vinluan, A. A. (2020). Predicting River Pollution Using Random Forest Decision Tree with GIS Model: A Case Study of MMORS, Philippines. *International Journal of Environmental Science and Development*, *11*(1), 36–42. https://doi.org/10.18178/ijesd.2020.11.1.1222

Wolfram, J., Stehle, S., Bub, S., Petschick, L. L. & Schulz, R. (2021). Water quality and ecological risks in European surface waters – Monitoring improves while water quality decreases. *Environment International*, *152*, 106479. https://doi.org/10.1016/j.envint.2021.106479

Wright, V. (2019). *Machine Learning: Using the Logistic Regression Model to Predict Coronary Heart Disease*. https://www.wrightanalytics-mn.com/pages/Machine_Learning_Using_the_Logistic_Regression_Model_to_Predict_Coronary_Heart_Final.pdf

Zhang, Z., Zhao, Y., Canes, A., Steinberg, D. & Lyashevska, O. (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*, *7*(7), 152–152. https://doi.org/10.21037/atm.2019.03.29