Department of Digital Business
**Journal of Artificial Intelligence and Digital Business (RIGGS)**
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 1 (2026) pp: 3958-3965
P-ISSN: 2963-9298, e-ISSN: 2963-914X

# Implementasi Algoritma Random Forest Classifier Dalam Klasifikasi Kelayakan Air Minum

Aulia Puspa [1], Aviarini Indrati[2]
[1,2]Sistem Informasi, Fakultas Ilmu Komputer Universitas Gunadarma
[1]auliapuspa27@gmail.com, [2]avi@staff.gunadarma.ac.id

***Abstrak***

*Air merupakan komponen esensial bagi tubuh manusia karena berperan penting dalam menjaga keseimbangan fisiologis, metabolisme, serta fungsi organ vital. Kualitas air minum yang tidak memenuhi standar kelayakan dapat berdampak negatif terhadap kesehatan, sehingga diperlukan metode yang efektif untuk mengklasifikasikan kelayakan air minum secara akurat. Salah satu pendekatan yang dapat digunakan adalah machine learning, khususnya algoritma Random Forest Classifier, yang mampu menganalisis pola data dan menghasilkan prediksi yang andal. Penelitian ini bertujuan untuk membangun model klasifikasi kelayakan air minum menggunakan algoritma Random Forest Classifier serta mengevaluasi tingkat akurasi yang dihasilkan oleh model tersebut. Dataset yang digunakan adalah drinking water potability yang diperoleh dari situs Kaggle, yang terdiri dari berbagai parameter kualitas air. Tahapan penelitian meliputi pengumpulan data, praproses data, pemodelan, dan evaluasi model. Proses praproses mencakup penanganan missing value, seleksi fitur, serta pembagian data menjadi data latih dan data uji. Pemodelan dilakukan menggunakan bahasa pemrograman Python dengan bantuan platform Google Colaboratory. Evaluasi model dilakukan menggunakan confusion matrix dan classification report untuk mengukur performa model berdasarkan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma Random Forest Classifier mampu mengklasifikasikan kelayakan air minum dengan tingkat akurasi sebesar 82%, sehingga model ini dapat digunakan sebagai metode yang efektif dalam mendukung proses klasifikasi kualitas air minum secara otomatis dan berbasis data.*

***Kata kunci**: Machine Learning, Random Forest Classifier, Air Minum.*

## 1. Latar Belakang

Air merupakan komponen esensial yang sangat penting bagi kehidupan manusia dan berperan dalam hampir seluruh proses biologis yang terjadi di dalam tubuh. Secara fisiologis, tubuh manusia sebagian besar tersusun atas air, yang berfungsi sebagai media transportasi nutrisi, pengatur suhu tubuh, serta pelarut berbagai zat kimia yang diperlukan untuk metabolisme sel. Kandungan air dalam tubuh manusia bervariasi tergantung pada jenis organ dan jaringan, dimana otak manusia tersusun sekitar 95% air, darah sekitar 75%, jantung sekitar 86%, dan ginjal sekitar 83% air (Kusumawardani & Larasati, 2020). Kandungan air yang tinggi pada organ-organ vital tersebut menunjukkan bahwa air memiliki peran fundamental dalam menjaga fungsi fisiologis tubuh secara optimal. Tanpa asupan air yang cukup dan berkualitas baik, berbagai fungsi tubuh dapat terganggu, termasuk sistem sirkulasi, metabolisme, ekskresi, dan regulasi suhu tubuh.

Selain berperan dalam menjaga fungsi fisiologis tubuh, air juga memiliki peran penting dalam menjaga kesehatan dan keseimbangan sistem biologis manusia. Air yang dikonsumsi harus memenuhi standar kualitas tertentu agar aman bagi kesehatan. Air minum yang layak konsumsi harus bebas dari kontaminan biologis, kimia, maupun fisik yang dapat membahayakan kesehatan manusia. Namun, pada kenyataannya masih banyak sumber air yang terkontaminasi oleh berbagai zat berbahaya seperti bakteri, virus, logam berat, dan bahan kimia lainnya yang dapat menyebabkan berbagai penyakit serius. Konsumsi air yang tidak aman dapat menyebabkan berbagai gangguan kesehatan, mulai dari penyakit ringan hingga penyakit kronis yang mengancam jiwa (World Health Organization, 2019).

Organisasi Kesehatan Dunia (World Health Organization) melaporkan bahwa air minum yang tidak aman merupakan salah satu penyebab utama berbagai penyakit menular, terutama di negara berkembang. Diperkirakan sekitar 829.000 orang meninggal setiap tahunnya akibat penyakit diare yang disebabkan oleh konsumsi air minum yang terkontaminasi (World Health Organization, 2019). Selain itu, pada tahun 2017 lebih dari 220 juta orang di seluruh dunia membutuhkan pengobatan pencegahan untuk schistosomiasis, yaitu penyakit kronis yang

disebabkan oleh parasit yang berkembang di lingkungan perairan yang terkontaminasi (World Health Organization, 2019). Data tersebut menunjukkan bahwa kualitas air minum merupakan faktor penting yang secara langsung mempengaruhi kesehatan masyarakat global. Oleh karena itu, pemantauan dan pengendalian kualitas air minum menjadi sangat penting untuk mencegah berbagai penyakit yang disebabkan oleh air yang tidak layak konsumsi.

Kualitas air minum dapat ditentukan berdasarkan berbagai parameter fisik, kimia, dan biologis. Parameter fisik meliputi kekeruhan (turbidity), warna, suhu, dan total padatan terlarut (total dissolved solids). Parameter kimia meliputi pH, kandungan logam berat, kadar klorin, sulfat, dan berbagai senyawa kimia lainnya. Sedangkan parameter biologis meliputi keberadaan mikroorganisme patogen seperti bakteri dan virus. Parameter-parameter tersebut dapat digunakan untuk menentukan apakah air tersebut layak atau tidak layak untuk dikonsumsi manusia (Ahmed & Mumtaz, 2019). Analisis terhadap parameter-parameter tersebut sangat penting untuk memastikan bahwa air yang dikonsumsi memenuhi standar kualitas yang telah ditetapkan oleh otoritas kesehatan.

Seiring dengan perkembangan teknologi informasi dan komputasi, metode analisis kualitas air minum telah mengalami perkembangan yang signifikan. Salah satu pendekatan yang saat ini banyak digunakan adalah metode berbasis machine learning. Machine learning merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk mempelajari pola dari data dan membuat prediksi berdasarkan pola tersebut tanpa harus diprogram secara eksplisit (Chen & Zhang, 2020). Metode machine learning dapat digunakan untuk menganalisis data kualitas air dan mengklasifikasikan apakah air tersebut layak atau tidak layak untuk dikonsumsi berdasarkan parameter-parameter tertentu.

Penggunaan machine learning dalam analisis kualitas air memiliki beberapa keunggulan dibandingkan metode konvensional. Metode machine learning mampu mengolah data dalam jumlah besar dengan tingkat akurasi yang tinggi dan waktu yang relatif singkat. Selain itu, metode ini juga mampu mengidentifikasi pola yang kompleks dan hubungan non-linear antar variabel yang sulit dideteksi menggunakan metode statistik tradisional (Kumar & Sharma, 2020). Oleh karena itu, machine learning menjadi salah satu metode yang efektif untuk digunakan dalam analisis kualitas air minum.

Salah satu algoritma machine learning yang banyak digunakan dalam klasifikasi kualitas air adalah Random Forest Classifier. Random Forest merupakan algoritma berbasis ensemble learning yang bekerja dengan membangun banyak pohon keputusan (decision tree) dan menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan prediksi akhir yang lebih akurat (Singh & Gupta, 2020). Algoritma ini memiliki keunggulan dalam menangani data dengan jumlah variabel yang banyak serta mampu mengurangi risiko overfitting yang sering terjadi pada algoritma decision tree tunggal. Selain itu, Random Forest juga memiliki kemampuan untuk menentukan tingkat kepentingan setiap fitur dalam proses klasifikasi, sehingga dapat membantu dalam mengidentifikasi parameter kualitas air yang paling berpengaruh terhadap kelayakan air minum (Shukla & Tiwari, 2020).

Beberapa penelitian sebelumnya telah menunjukkan bahwa algoritma Random Forest memiliki kinerja yang sangat baik dalam klasifikasi kualitas air minum. Penelitian yang dilakukan oleh Devi (2019) menunjukkan bahwa algoritma Random Forest mampu menghasilkan tingkat akurasi sebesar 93,75% dalam memprediksi kualitas air berdasarkan berbagai parameter kualitas air. Penelitian tersebut juga menemukan bahwa parameter Total Dissolved Solids (TDS) merupakan salah satu variabel yang paling signifikan dalam menentukan kualitas air. Selain itu, penelitian lain yang dilakukan oleh Singh (2021) menunjukkan bahwa algoritma Random Forest mampu mencapai tingkat akurasi sebesar 97% dalam klasifikasi kualitas air perkotaan, yang lebih tinggi dibandingkan algoritma machine learning lainnya seperti Support Vector Machine dan Decision Tree.

Penelitian lain yang dilakukan oleh Riyantoko et al. (2021) juga menunjukkan bahwa algoritma Random Forest mampu menghasilkan tingkat akurasi sebesar 72,81% dalam klasifikasi kualitas air minum berdasarkan parameter kualitas air yang tersedia dalam database. Hasil penelitian tersebut menunjukkan bahwa algoritma Random Forest memiliki kemampuan yang baik dalam mengklasifikasikan kualitas air minum, meskipun tingkat akurasi yang dihasilkan dapat dipengaruhi oleh jumlah dan kualitas data yang digunakan. Selain itu, penelitian lain juga menunjukkan bahwa metode machine learning, khususnya Random Forest, memiliki tingkat akurasi yang tinggi dan stabil dalam memprediksi kualitas air dibandingkan metode konvensional (Iqbal & Khan, 2019).

Keunggulan lain dari algoritma Random Forest adalah kemampuannya dalam menangani data yang memiliki missing value dan data yang tidak seimbang. Dalam banyak kasus, dataset kualitas air seringkali memiliki data yang tidak lengkap atau tidak seimbang, yang dapat mempengaruhi kinerja model klasifikasi. Random Forest memiliki mekanisme internal yang memungkinkan algoritma ini tetap dapat menghasilkan prediksi yang akurat meskipun terdapat data yang tidak lengkap (Patel & Jain, 2020). Selain itu, algoritma ini juga mampu mengurangi

DOI: https://doi.org/10.31004/riggs.v5i1.6773

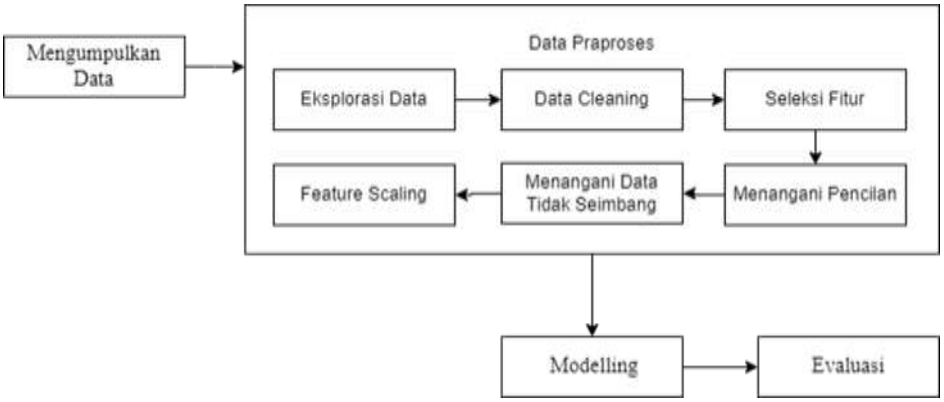
variansi model dan meningkatkan stabilitas prediksi melalui penggunaan teknik ensemble learning (Verma & Singh, 2020).

Mengingat pentingnya kualitas air minum bagi kesehatan manusia serta kemampuan algoritma Random Forest dalam mengklasifikasikan kualitas air dengan tingkat akurasi yang tinggi, maka penelitian ini bertujuan untuk membangun model klasifikasi kelayakan air minum menggunakan algoritma Random Forest Classifier. Model yang dibangun diharapkan dapat digunakan untuk mengklasifikasikan kelayakan air minum berdasarkan parameter kualitas air yang tersedia dalam dataset. Dengan adanya model klasifikasi ini, diharapkan proses identifikasi kelayakan air minum dapat dilakukan secara lebih cepat, akurat, dan efisien dibandingkan metode konvensional.

Selain itu, penelitian ini juga diharapkan dapat memberikan kontribusi dalam pengembangan metode berbasis machine learning untuk analisis kualitas air minum. Penggunaan metode machine learning dalam analisis kualitas air dapat membantu dalam meningkatkan efektivitas sistem pemantauan kualitas air serta mendukung upaya pencegahan penyakit yang disebabkan oleh konsumsi air yang tidak layak. Dengan demikian, penelitian ini memiliki potensi untuk memberikan manfaat tidak hanya dalam bidang teknologi informasi, tetapi juga dalam bidang kesehatan masyarakat dan pengelolaan sumber daya air secara berkelanjutan.

## 2. Metode Penelitian

Metode penelitian yang digunakan pada penelitian ini terdiri dari beberapa tahapan, yaitu pengumpulan data, praproses data, modeling dan evaluasi. Metode penelitian yang digunakan ditunjukkan pada Gambar 1:




Gambar 1 Metode Penelitian

### 1. Pengumpulan Dataset

Tahap pertama dalam penelitian ini adalah pengumpulan dataset yang digunakan sebagai bahan utama dalam proses analisis dan pemodelan. Dataset yang digunakan adalah dataset *drinking water potability* yang diperoleh dari situs Kaggle, yaitu sebuah platform penyedia dataset terbuka yang sering digunakan dalam penelitian data science dan machine learning. Dataset tersebut berisi informasi mengenai berbagai parameter kualitas air yang berpengaruh terhadap kelayakan air minum, seperti pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, dan turbidity, serta satu variabel target yaitu potability yang menunjukkan apakah air tersebut layak diminum atau tidak. Dataset yang telah diunduh kemudian disimpan dalam format comma separated value (CSV), karena format ini mudah dibaca dan diproses menggunakan berbagai perangkat lunak pengolahan data, termasuk bahasa pemrograman Python.

### 2. Data Praproses

Tahap selanjutnya adalah praproses data, yang bertujuan untuk membersihkan dan mempersiapkan data mentah agar dapat digunakan secara optimal dalam proses pemodelan. Praproses data dilakukan menggunakan bahasa pemrograman Python dengan bantuan platform Google Colaboratory, yang menyediakan lingkungan pemrograman berbasis cloud. Tahapan praproses dimulai dengan eksplorasi data (exploratory data analysis) untuk memahami karakteristik dataset, termasuk jumlah data, jenis variabel, distribusi data, serta keberadaan nilai yang hilang (missing value). Selanjutnya dilakukan proses data cleaning untuk menangani missing value dengan metode imputasi, sehingga tidak ada data yang kosong. Setelah itu dilakukan seleksi fitur untuk menentukan variabel yang paling relevan dalam memprediksi kelayakan air minum. Penanganan pencilan (outlier) juga dilakukan untuk mengurangi pengaruh data ekstrem yang dapat menurunkan performa model. Selain itu, dilakukan penanganan data tidak seimbang (imbalanced data) menggunakan teknik undersampling NearMiss, yang bertujuan untuk menyeimbangkan jumlah data antara kelas layak minum dan tidak layak minum. Tahap terakhir adalah feature

DOI: https://doi.org/10.31004/riggs.v5i1.6773

scaling, yaitu proses normalisasi data agar seluruh variabel memiliki skala yang sama, sehingga model dapat bekerja lebih optimal dan tidak bias terhadap variabel tertentu.

**3. Modeling**

Tahap modeling merupakan proses pembangunan model klasifikasi menggunakan algoritma Random Forest Classifier. Sebelum model dibangun, dataset terlebih dahulu dibagi menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data), dengan proporsi 80% untuk data latih dan 20% untuk data uji. Data latih digunakan untuk melatih model agar dapat mengenali pola dalam data, sedangkan data uji digunakan untuk mengukur kemampuan model dalam memprediksi data yang belum pernah dilihat sebelumnya. Model Random Forest kemudian diinisialisasi dengan parameter n_estimators = 100, yang berarti model menggunakan 100 pohon keputusan, criterion = 'entropy' untuk mengukur kualitas pemisahan data, serta random_state = 0 untuk memastikan hasil yang konsisten. Setelah parameter ditentukan, model dilatih menggunakan data latih melalui proses fitting. Hasil dari proses ini adalah model yang mampu mengklasifikasikan kelayakan air minum berdasarkan parameter kualitas air yang tersedia.

**4. Evaluasi**

Tahap terakhir adalah evaluasi model, yang bertujuan untuk mengukur performa model dalam melakukan klasifikasi. Evaluasi dilakukan menggunakan confusion matrix dan classification report. Confusion matrix digunakan untuk membandingkan hasil prediksi model dengan nilai aktual, sehingga dapat diketahui jumlah prediksi yang benar dan salah, termasuk true positive, true negative, false positive, dan false negative. Sementara itu, classification report digunakan untuk menampilkan metrik evaluasi yang lebih rinci, yaitu accuracy, precision, recall, dan F1-score. Accuracy menunjukkan tingkat ketepatan model secara keseluruhan, precision menunjukkan tingkat ketepatan prediksi positif, recall menunjukkan kemampuan model dalam mendeteksi kelas positif, dan F1-score merupakan kombinasi dari precision dan recall. Hasil evaluasi ini digunakan untuk menilai apakah model yang dibangun memiliki performa yang baik dalam mengklasifikasikan kelayakan air minum.

## 3. Hasil dan Diskusi

**1. Dataset**

Dataset yang digunakan dalam penelitian ini adalah dataset *drinking water potability* yang diperoleh dari platform Kaggle. Dataset ini terdiri dari 3276 baris data dan 10 kolom, yang mencakup 9 variabel fitur (feature variables) dan 1 variabel target (target variable). Variabel target dalam dataset ini adalah *Potability*, yang menunjukkan apakah air tersebut layak untuk dikonsumsi manusia atau tidak. Variabel ini memiliki dua kategori, yaitu nilai 1 yang menunjukkan air layak minum dan nilai 0 yang menunjukkan air tidak layak minum.

Dataset ini digunakan secara luas dalam penelitian klasifikasi kualitas air karena mencakup parameter penting yang berhubungan langsung dengan kelayakan air minum. Menurut World Health Organization, kualitas air minum harus memenuhi standar tertentu untuk mencegah risiko kesehatan seperti penyakit diare, infeksi parasit, dan gangguan organ tubuh (World Health Organization, 2019). Oleh karena itu, parameter kualitas air seperti pH, total dissolved solids, dan turbidity menjadi indikator penting dalam menentukan kelayakan air minum.

1. **pH** merupakan indikator tingkat keasaman atau kebasaan air. Nilai pH air minum yang aman umumnya berada dalam rentang 6,5 hingga 8,5. Air dengan pH yang terlalu rendah bersifat asam dan dapat menyebabkan korosi pada sistem distribusi air, sedangkan air dengan pH yang terlalu tinggi dapat menyebabkan gangguan kesehatan dan rasa tidak enak (WHO, 2019).
2. **Hardness** menunjukkan tingkat kesadahan air yang disebabkan oleh kandungan kalsium dan magnesium. Air dengan tingkat kesadahan yang tinggi dapat menyebabkan pembentukan kerak pada peralatan dan dapat mempengaruhi kualitas air minum (Kumar & Sharma, 2020).
3. **Total Dissolved Solids (TDS)** menunjukkan jumlah total zat padat terlarut dalam air, termasuk mineral, garam, dan logam. Nilai TDS yang tinggi dapat menunjukkan adanya kontaminasi dan mempengaruhi rasa serta keamanan air (Chen & Zhang, 2020).
4. **Chloramines** merupakan senyawa kimia yang digunakan sebagai disinfektan dalam sistem pengolahan air untuk membunuh mikroorganisme berbahaya. Namun, kadar chloramines yang berlebihan dapat berdampak negatif pada kesehatan manusia (Ahmed & Mumtaz, 2019).
5. **Sulfate** adalah mineral alami yang ditemukan dalam air. Kandungan sulfate yang tinggi dapat menyebabkan gangguan pencernaan dan mempengaruhi rasa air (Patel & Jain, 2020).
6. **Conductivity** mengukur kemampuan air dalam menghantarkan arus listrik, yang berkaitan dengan jumlah ion terlarut dalam air. Semakin tinggi nilai conductivity, semakin tinggi tingkat kontaminasi air (Das & Roy, 2019).

DOI: https://doi.org/10.31004/riggs.v5i1.6773

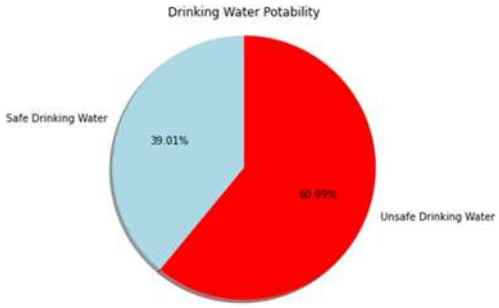
7. **Organic Carbon** menunjukkan jumlah karbon organik dalam air, yang dapat berasal dari bahan organik alami atau kontaminasi lingkungan (Shukla & Tiwari, 2020).
8. **Trihalomethanes** merupakan senyawa kimia yang terbentuk sebagai produk sampingan dari proses desinfeksi menggunakan klorin. Senyawa ini berpotensi menyebabkan risiko kesehatan jika dikonsumsi dalam jumlah tinggi (Singh & Gupta, 2020).
9. **Turbidity** menunjukkan tingkat kekeruhan air yang disebabkan oleh partikel tersuspensi. Air dengan tingkat kekeruhan tinggi dapat menunjukkan adanya kontaminasi mikroorganisme (Verma & Singh, 2020).
10. **Potability** merupakan variabel target yang menunjukkan apakah air layak dikonsumsi atau tidak.

Dataset ini memberikan dasar yang kuat untuk melakukan analisis klasifikasi menggunakan algoritma machine learning, khususnya Random Forest Classifier, karena mencakup parameter yang relevan dan signifikan dalam menentukan kualitas air minum.

**2. Data Bersih**

Tahap praproses data merupakan tahap penting dalam penelitian machine learning karena kualitas data sangat mempengaruhi performa model yang dihasilkan. Data yang bersih dan terstruktur dengan baik akan menghasilkan model dengan akurasi yang lebih tinggi dibandingkan data yang tidak bersih (Chen & Zhang, 2020).

Proporsi keseimbangan label dalam dataset divisualisasikan menggunakan pie chart.




Gambar 2 Proporsi Label

Berdasarkan Gambar 2, dapat diketahui bahwa proporsi data tidak seimbang. Data dengan label 0 (tidak layak minum) berjumlah 60,99%, sedangkan data dengan label 1 (layak minum) berjumlah 39,01%. Ketidakseimbangan data ini dapat menyebabkan model menjadi bias terhadap kelas mayoritas, sehingga perlu dilakukan teknik penyeimbangan data (imbalance handling) (Ahmed & Mumtaz, 2019).

Selain itu, ditemukan adanya missing value pada beberapa variabel, yaitu:

- pH : 491 missing value
- Sulfate : 781 missing value
- Trihalomethanes : 162 missing value

Missing value merupakan masalah umum dalam dataset dunia nyata dan dapat mempengaruhi performa model jika tidak ditangani dengan benar (Shukla & Tiwari, 2020). Oleh karena itu, dilakukan imputasi menggunakan nilai rata-rata (mean imputation). Metode ini dipilih karena mampu mempertahankan distribusi data tanpa mengurangi jumlah data secara signifikan.

Selanjutnya dilakukan analisis feature importance menggunakan algoritma Random Forest. Hasil menunjukkan bahwa fitur sulfate memiliki tingkat kepentingan tertinggi dengan nilai 0.15775261, sedangkan fitur conductivity, organic carbon, turbidity, dan trihalomethanes memiliki nilai kepentingan lebih rendah.

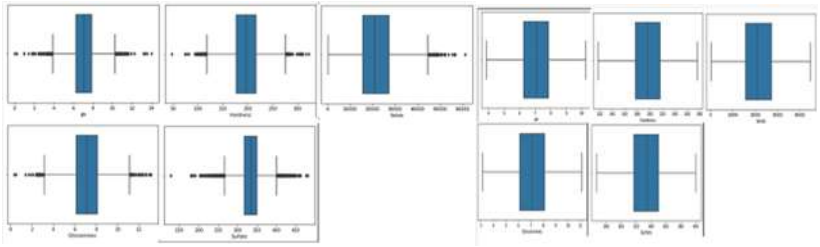
Oleh karena itu, dipilih lima fitur utama untuk proses modeling, yaitu:

- Chloramines
- Solids
- Hardness
- pH
- Sulfate

DOI: https://doi.org/10.31004/riggs.v5i1.6773

Pemilihan fitur yang relevan dapat meningkatkan performa model dan mengurangi kompleksitas komputasi (Singh, 2021).

Selanjutnya dilakukan analisis outlier menggunakan boxplot.



Gambar 3 Hasil Visualisasi Penanganan Pencilan

Boxplot menunjukkan distribusi data dan keberadaan outlier. Outlier dapat mempengaruhi performa model karena dapat menyebabkan bias dalam proses pembelajaran (Das & Roy, 2019).

Selain itu, dilakukan teknik undersampling menggunakan metode NearMiss untuk menyeimbangkan data. Hasilnya, jumlah data pada masing-masing kelas menjadi seimbang, yaitu:

- 1278 data layak minum
- 1278 data tidak layak minum

Selanjutnya dilakukan feature scaling menggunakan normalisasi Min-Max untuk mengubah nilai fitur ke rentang 0 hingga 1. Normalisasi ini penting untuk meningkatkan stabilitas dan performa model machine learning (Patel & Jain, 2020).

**3. Model Random Forest Classifier**

Random Forest Classifier digunakan dalam penelitian ini karena merupakan algoritma yang memiliki performa tinggi dalam klasifikasi dan mampu mengatasi overfitting dengan baik (Breiman, 2001).

Random Forest bekerja dengan membangun banyak pohon keputusan (decision trees) dan menggabungkan hasilnya untuk menghasilkan prediksi akhir.

```
RandomForestClassifier(criterion='entropy', random_state=0)
```

Gambar 4 Model Random Forest Classifer

Model dibangun menggunakan parameter:

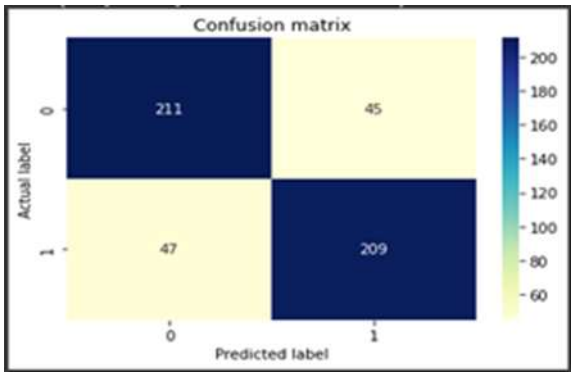
- n_estimators = 100
- criterion = entropy
- random_state = 0

Penggunaan banyak pohon keputusan meningkatkan akurasi model dan stabilitas prediksi (Chen & Zhang, 2020).

**4. Confusion Matrix**

Confuison matrix digunakan untuk membandingkan antara hasil prediksi yang dihasilkan oleh model random forest dengan actual value. Gambar dibawah ini menunjukkan hasil prediksi oleh model dimana terdapat 211 data True Negatif (TN), yaitu data-data yang sebenarnya bernilai negatif dan diprediksi benar negatif oleh model. Terdapat 209 data True Positif (TP), yaitu data yang sebenarnya bernilai positif dan di prediksi benar positif oleh model. Terdapat 45 data False Positif (FP), yaitu data yang sebenarnya bernilai negatif tetapi diprediksi positif oleh model dan 47 data False negative (FN), yaitu data yang sebenarnya bernilai positif tetapi diprediksi negatif oleh model.

DOI: https://doi.org/10.31004/riggs.v5i1.6773




Gambar 5 Confusion Matrix

**5. Classification Report**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.82 | 0.82 | 256 |
| 1 | 0.82 | 0.82 | 0.82 | 256 |
| accuracy | | | 0.82 | 512 |
| macro avg | 0.82 | 0.82 | 0.82 | 512 |
| weighted avg | 0.82 | 0.82 | 0.82 | 512 |

Gambar 6

Nilai accuracy yang diperoleh yaitu sebesar 82%. Nilai accuracy menunjukkan seberapa akurat model memprediksi dengan benar atau tingkat kedekatan nilai prediksi dan nilai aktual. Akurasi ini menggambarkan sebesar 82 persen air diprediksi layak maupun tidak dari keseluruhan air. Nilai ini belum menunjukkan hasil yang cukup baik untuk kasus klasifikasi air minum, karena ada kemungkinan 18% air yang terkontaminasi mengalami kesalahan prediksi yang berarti dapat terjadi kasus konsumsi air yang terkontaminasi zat berbahaya sebesar 18%.

Nilai Precision yang diperoleh adalah sebesar 82%. Nilai ini menunjukkan tingkat keakuratan antara data yang diminta dengan hasil prediksi model. Precision ini menggambarkan 82 persen akurat air yang sebenarnya layak minum dari keseluruhan air yang diprediksi layak minum. Nilai ini belum cukup baik karena ada kemungkinan 18% terjadi false positif, dimana air yang sebenarnya tidak layak, tetapi diprediksi layak minum.

Nilai Recall yang diperoleh adalah sebesar 82%. Nilai recall atau peluang kasus dengan kategori positif dan tepat. Nilai recall menggambarkan 82 persen akurat air yang diprediksi layak minum dibandingkan keseluruhan air yang sebenarnya layak minum. Nilai ini belum cukup baik, karena ada kemungkinan 18% terjadi false negatif, dimana algoritma memprediksi air tidak layak minum, padahal sebenarnya air tersebut layak minum.

Nilai F1-Score menghasilkan skor sebesar 82%. Nilai ini mengambil nilai presisi dan recall untuk memberikan representasi standar dari nilai – nilai tersebut.

## 4. Kesimpulan

Model klasifikasi Random Forest untuk mengklasifikasikan kelayakan air minum telah berhasil dikembangkan dan diimplementasikan dengan baik melalui tahapan pengumpulan data, praproses data, pemodelan, serta evaluasi performa model. Dataset yang digunakan dalam penelitian ini merupakan dataset drinking water potability yang diperoleh dari platform Kaggle, yang berisi berbagai parameter kualitas air yang relevan dalam menentukan kelayakan air untuk dikonsumsi. Pada tahap praproses data, dilakukan beberapa proses penting, seperti penanganan missing value, seleksi fitur, penanganan ketidakseimbangan data menggunakan metode undersampling NearMiss, serta feature scaling untuk memastikan data berada dalam rentang yang seragam. Tahapan ini bertujuan untuk meningkatkan kualitas data sehingga model dapat bekerja secara optimal dan menghasilkan prediksi yang lebih akurat. Dalam penelitian ini, proses seleksi fitur dilakukan dengan mempertimbangkan nilai feature importance yang dihasilkan oleh algoritma Random Forest. Berdasarkan hasil seleksi tersebut, diperoleh lima fitur yang memiliki tingkat kepentingan paling tinggi terhadap variabel target, yaitu kloramin (chloramines), padatan terlarut (solids), kesadahan (hardness), pH, dan sulfat (sulfate). Pemilihan fitur-fitur tersebut terbukti mampu meningkatkan efisiensi model dengan mengurangi kompleksitas tanpa mengurangi kemampuan prediksi secara signifikan. Dengan menggunakan lima fitur utama tersebut, model Random Forest berhasil mencapai nilai akurasi sebesar 82%, yang menunjukkan bahwa model mampu mengklasifikasikan data air layak minum dan tidak layak minum dengan tingkat ketepatan yang cukup baik. Selain nilai akurasi, hasil evaluasi model juga menunjukkan

DOI: https://doi.org/10.31004/riggs.v5i1.6773

nilai precision, recall, dan f1-score yang seimbang, yang mengindikasikan bahwa model memiliki kemampuan yang konsisten dalam mengidentifikasi kedua kelas secara proporsional. Hal ini menunjukkan bahwa algoritma Random Forest merupakan metode yang efektif dan andal dalam mengklasifikasikan kelayakan air minum berdasarkan parameter kualitas air. Dengan demikian, model yang dikembangkan dalam penelitian ini dapat dijadikan sebagai dasar dalam pengembangan sistem pendukung keputusan untuk membantu proses identifikasi dan pemantauan kualitas air minum secara otomatis dan efisien di masa mendatang.

## Referensi

1. Ahmed, U., & Mumtaz, R. (2019). Prediction of water quality using machine learning algorithms. *Journal of Environmental Management, 234*, 256–264. https://doi.org/10.1016/j.jenvman.2018.11.049
2. Chen, H., & Zhang, Y. (2020). Machine learning for water quality monitoring: A review. *Environmental Science and Pollution Research, 27*, 43990–44003.
3. Das, T. K., & Roy, S. (2019). Water quality prediction using artificial intelligence techniques. *Journal of Water Resource and Protection, 11*(8), 1012–1023.
4. Devi, G. (2019). Random forest advice for water quality prediction in the regions of Kadapa District. *International Journal of Innovative Technology and Exploring Engineering, 8*, 1464–1466. https://doi.org/10.35940/ijitee.F1298.0486S419
5. Iqbal, M. H., & Khan, A. (2019). Water quality prediction using supervised machine learning techniques. *International Journal of Advanced Computer Science and Applications, 10*(7), 1–7.
6. Kavitha, S., & Saravanan, R. (2020). Water quality classification using machine learning algorithms. *International Journal of Scientific and Technology Research, 9*(3), 4210–4215.
7. Kumar, P., & Sharma, A. (2020). Prediction of drinking water quality using random forest and support vector machine. *Journal of Environmental Informatics, 35*(2), 145–156.
8. Kusumawardani, S., & Larasati, A. (2020). Analisis konsumsi air putih terhadap konsentrasi siswa. *Holistika: Jurnal Ilmiah PGSD, 4*(2), 91–95.
9. Patel, N., & Jain, S. (2020). Machine learning approaches for water quality analysis: A comparative study. *International Journal of Computer Applications, 177*(22), 15–21.
10. Riyantoko, P. A., Fahrudin, T. M., & Hindrayani, K. M. (2021). Analisis sederhana pada kualitas air minum berdasarkan akurasi model klasifikasi dengan menggunakan Lucifer machine learning. *Seminar Nasional Sains Data, 1*, 12–18.
11. Shukla, S. K., & Tiwari, P. K. (2020). Water quality assessment using machine learning and artificial intelligence. *Environmental Monitoring and Assessment, 192*(7), 1–12.
12. Singh, B. J. (2021). Smart urban water quality prediction system using machine learning. *Journal of Physics: Conference Series, 1979*(1), 012057. https://doi.org/10.1088/1742-6596/1979/1/012057
13. Singh, R. K., & Gupta, M. (2020). Application of random forest algorithm in water quality prediction. *International Journal of Engineering Research and Technology, 9*(5), 234–239.
14. Verma, A., & Singh, S. (2020). Assessment of drinking water quality using data mining techniques. *International Journal of Computer Science and Information Security, 18*(4), 45–52.
15. World Health Organization. (2019). Drinking water. https://www.who.int/news-room/fact-sheets/detail/drinking-water

---

DOI: https://doi.org/10.31004/riggs.v5i1.6773