Analisis Kualitas Air PAM Layak Minum dengan Metode *Random Forest* dan *Decision Tree*

Stefani Kelin Martha Ampak[a1], AAIN Eka Karyawati[a2], I Komang Arya Ganda Wiguna[a3]

[a]Program Studi Informatika, Universitas Udayana
Kuta Selatan,badung,Bali, Indonesia
[1]stefanikelin6@email.com
[2]eka.karyawati@unud.ac.id
arya.ganda@unud.ac.id

Abstract

Water is an important source of life for living things including humans. Human needs for water include water that is suitable for use in cooking, washing, and bathing activities that are clean and healthy, as well as water that is safe to drink. Drinking Water Companies (PAM) have a vital role in providing water that meets the standards of consumption eligibility. This study aims to analyze the quality of PAM water by utilizing the Random Forest method as a classification method. The data used includes physical, chemical, and microbiological parameters of water. The use of the random forest method was chosen because of its ability to handle complex data and produce accurate predictions. The results of the study showed that the random forest model was able to classify water quality with a high level of accuracy and identify the parameters that most influence the eligibility of drinking water. This study is expected to help related parties in monitoring and improving the quality of PAM water so that it is in accordance with the established health standards.

Keywords: *water quality, PAM, drinkable, Random Forest, classification*

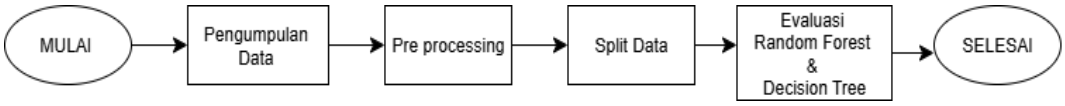
1. Pendahuluan

Kehidupan manusia bergantung pada air untuk segala hal. namun, berbagai masalah lingkungan dan kesehatan dapat terjadi akibat kualitas air yang buruk [1] Berdasarkan data *World Health Organization* (WHO) sebanyak 1,8 miliar penduduk dunia akan menghadapi ‘kelangkaan air mutlak’ yaitu tidak bisa memenuhi kebutuhan air minimal 500-meter kubik per tahun per kapita. Kebutuhan akan air bersih di Indonesia terus meningkat, namun aksesibilitas nya masih terbatas. Hal ini terjadi di akibatkan pembangunan yang tidak memperhatikan keseimbangan wilayah dan mengurangi daerah resapan, terutama di daerah perkotaan. akibatnya, tidak banyak sumber air bersih yang tersedia. ketersediaan air yang layak minum tidak hanya berpengaruh terhadap kesehatan, tetapi juga terhadap kualitas hidup masyarakat secara keseluruhan. di Indonesia, salah satu penyedia utama air bersih adalah Perusahaan Air Minum (PAM). perusahaan Air Minum (PAM) adalah air yang berasal dari sungai dan telah mengalami proses penjernihan. Namun, dalam praktiknya masih di temukan kasus air PAM yang tidak memenuhi standar kelayakan untuk dikonsumsi, baik karena kontaminasi fisik, kimia maupun mikrobiologis. Menurut *Advanced Analytics Asia*, air yang terkontaminasi dapat menyebabkan berbagai penyakit pada manusia seperti: keracunan, gangguan pencernaan, infeksi saluran pernapasan, dan berbagai masalah kesehatan lainnya. peraturan menteri kesehatan terdapat parameter yang digunakan untuk menentukan kualitas air seperti keasaman (pH), Sulfat, Logam terlarut, kekeruhan, warna, dan total *coliform*. untuk memastikan air PAM aman dikonsumsi, diperlukan proses evaluasi yang akurat dan efisien. keberadaan zat-zat yang terdapat di dalam air dapat menjadi indikator untuk menentukan kualitas air. pada penilaian kualitas air menggunakan metode perhitungan manual seperti *Water Quality Index* (WQI). *Water Quality Index* adalah metode sederhana yang digunakan sebagai bagian dari survei kualitas air secara umum dengan menggunakan sekelompok parameter yang mengurangi sejumlah besar informasi ke nomor tunggal, biasanya berdimensi, dengan cara yang mudah di reproduksi. Metode konvensional bergantung pada pemahaman terhadap parameter-parameter yang telah ditetapkan untuk mengevaluasi dan

mengkategorikan kualitas air sebagai air layak di minum. Metode konvensional sering kali memerlukan waktu lama dalam perhitungan nya, sehingga membutuhkan sistem yang dapat bekerja secara otomatis [2]. Penelitian ini menggunakan *dataset Water Quality Indeks*, berisi data kualitas air dengan *value* yang berbeda-beda pada 10 Atribut yang di miliki. *attribute potability* merupakan *attributte* yang diberi label sebagai hasil *conclusion* apakah air tersebut layak di minum atau tidak untuk digunakan berdasarkan hasil akurasi dari masing-masing-masing *value attribute* yang ada pada data tersebut. Menurut penelitian [3] untuk memprediksi kelayakan air minum dengan menggunakan algoritma model *Random Forest*. Dalam penelitian nya di dapatkan hasil akurasi sebesar 69% dan penelitian nya mempertimbangkan penerapan algoritma *Random Forest* sebagai alat prediktif dalam penilaian kualitas air. Penelitian lain oleh [4] tujuan penelitian ini adalah untuk pengukuran untuk melihat tingkat keparahan penyakit pada daun apel menggunakan metode algoritma *Random Forest*, menunjukkan bahwa tingkat akurasi pada proses pengujian sebesar 75.3191%. Dalam penelitian [5] menegaskan bahwa algoritma *Decision Tree* memiliki keunggulan dalam memetakan keputusan klasifikasi kualitas air di bandingkan SVM dan *Logistic Regression*. Penelitian lain [6]dengan membandingkan metode algoritma *Decision Tree*, *Logistic Regression*, *Support Vector Machine* (SVM), dan *Artificiaal Neural Network* (ANN). Hasil akurasi metode penelitian tersebut adalah *Decision Tree* 60.19%, *Logistic Regression* 62.80%, *SVM* 68.59%, dan ANN 69.54%. Berdasarkan permasalahan tersebut, tujuan penelitian yang dilakukan untuk melakukan perbandingan dengan menggunakan kedua metode : *Random Forest* dan *Decision Tree* maka dilakukan penelitian dengan judul "Analisis Kualitas Air PAM Layak Minum Dengan Metode *Random Forest* dan *Decision Tree*" dengan menggunakan *tools Google Collab* untuk mengetahui performa keunggulan algoritma *Random Forest* dan *Decision Tree* dengan nilai akurasi yang paling besar dari kedua metode yang akan di implementasikan ke dalam klasifikasi data. Penelitian ini bertujuan untuk menganalisis kelayakan kualitas air PAM dengan menggunakan Algoritma *Random Forest* dan *Decision Tree*. diharapkan hasil penelitian ini dapat memberikan gambaran yang lebih objektif tentang kondisi air yang di salurkan kepada masyarakat serta menjadi referensi dalam pengambilan keputusan bagi pihak penyedia layanan air bersih.

2. Metode Penelitian

Metode penelitian adalah pendekatan sistematis yang digunakan oleh peneliti untuk mengumpulkan, menganalisis dan menginterpretasikan data. Penelitian ini merupakan penelitian eksperimen. Pengumpulan data dalam penelitian ini berasal dari *Water Quality Dataset Kaggle*, dimana terdapat 3276 *record*, 10 atribut dan 2 target kelas yang kemudian akan di olah menggunakan *tools Google Collab* dan data akan dibagi untuk data *training* dan data *testing* dengan *Python Programming m*enggunakan algoritma *Random Forest* dan *Decision Tree.* Berikut langkah-langkah untuk mendapatkan nilai akurasi pada penelitian ini adalah sebagai berikut:



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Penelitian ini menggunakan data sekunder dari *Water Quality Dataset* yang diambil melalui *website kaggle.com* dengan jumlah data yang diperoleh 3276 dan 10 atribut. Untuk proses klasifikasi dalam membandingkan hasil akurasi dari ketiga metode yang digunakan yaitu *Random Forest* dan *Decision Tree.* Berikut tabel 1 atribut yang digunakan dalam pengumpulan data.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Tabel 1. *Data set*

Atrribut	Deskripsi
ph	pH 1. air (0 hingga 14)
Hardness	Kapasitas air untuk mengendapkan sabun dalam mg/L
Solids	Total padatan terlarut dalam
Chloramines	Jumlah Kloramina dalam ppm
Sulfate	Jumlah Sulfat yang terlarut dalam mg/L
Conductivity	Konduktivitas listrik air dalam µS/cm.
Organic_Carbon	Jumlah karbon organik dalam ppm.
Trihalomethanes	Jumlah Trihalometan dalam µg/L.
Turbidity	Ukuran sifat air yang memancarkan cahaya dalam NTU.
Potability	Menunjukkan apakah air aman untuk dikonsumsi manusia. Dapat diminum - 1 dan Tidak dapat diminum -0

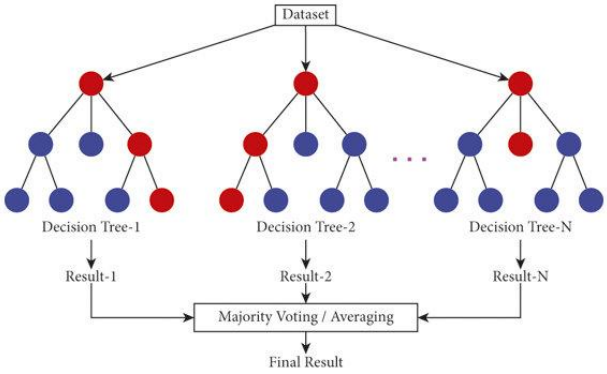
2.2 *Pre-Processing* & Split Data

Data yang didapatkan dalam penelitian ini perlu dilakukan nya proses *pre-processing. pre-processing* adalah tahapan untuk menghilangkan beberapa permasalahan yang bisa mengganggu saat pemrosesan data [7]. proses labeling data dalam penelitian ini yaitu *Potability. Potability* menjadi *attribute* untuk mengetahui apakah air aman layak diminum (1) atau tidak (0). beberapa *pre-processing* yang digunakan seperti resampling dilakukan pada *dataset* untuk menghilangkan masalah ketidakseimbangan data. kemudian menghilangkan *missing value* (nilai null), yang membuat perhitungan menjadi lebih mudah. pada penelitian ini setelah dilakukan proses *pre-processing* data maka data yang awalnya 3276 data menjadi 2.011 data. Tahap selanjutnya, melakukan tahap Split data untuk melakukan pembagian antara data pelatihan dan data pengujian. pembagian ini dilakukan dengan rasio 80% untuk data pelatihan dan 20% untuk data pengujian. maka, menghasilkan jumlah data sebesar 1.920 untuk data pelatihan (data *training*) dan sebesar 480 untuk data pengujian (data *testing*).

2.3 Implementasi *Random Forest & Decision Tree*

Tahap kunci dalam penelitian ini melibatkan pengembangan model klasifikasi dengan menerapkan algoritma *Random Forest* dan *Decision Tree*. Metode *Random Forest* ini menggabungkan beberapa pohon Keputusan untuk bekerja bersama-sama, menghasilkan yang menentukan hasil akhri dalam mendeteksi sarkasme [8]. Gambar 2 merupakan ilustrasi algoritma *random forest*.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 2. Ilustrasi algoritma *Random Forest*

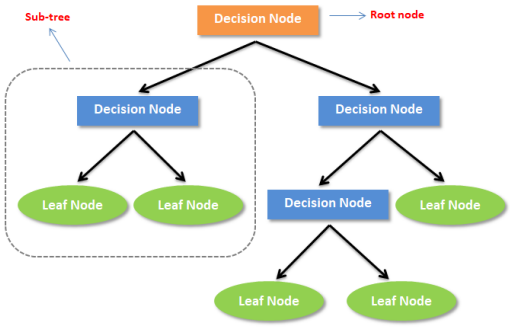
Persamaan yang diterapkan dalam algoritma *Random Forest* dapat di identifikasi pada rumus (1).

$$fi_i = = \frac{\sum j: node\ \ j\ splits\ on\ featre\ \ i\ ni_j}{\sum \epsilon all\ nodes\ \ ni_k} \times 100\% \quad (1)$$

Keterangan:

fi_i = fitur i
ni_j = sampul j

Metode Decision Tree ini memecahkan ke dalam kelompok yang lebih kecil berdasarkan atribut di dalam data [9]. Pembagian kelompok ini dilakukan berulang kali hingga seluruh elemen data yang berasal dari kelas yang sama dapat masuk ke dalam satu kelompok. Gambar 3 merupakan ilustrasi algoritma *decision tree*.



Gambar 3. Ilustrasi algoritma *Decision Tree*

Persamaan yang diterapkan dalam algoritma *Decision Tree* dapat di identifikasi pada rumus (2).

$$Entropy(S) = = -\text{Entropy} = \sum \text{n i} = 1\ \text{pi}\ *\ \log 2\ (\text{pi} * \log 2\ (\text{pi}) \quad (2)$$

Keterangan:

S = Himpunan kasus
n = jumlah partisi s
pi = jumlah kasus pada partisi ke-i

p-ISSN: 2986-3929
e-ISSN: 3032-1948

2.4 Evaluasi

Confusion Matrix adalah merupakan metode klasifikasi berdasarkan hasil klasifikasi yang telah di lakukan, dimana akurasi klasifikasi mempengaruhi kinerja klasifikasi [10]. pentingnya *confusion matrix* akan memberikan informasi beberapa baik model yang telah dibuat melalui pengukuran akurasi yang ada untuk mengetahui seberapa akurat model yang telah dibuat. *confusion matrix* digunakan untuk menghitung *accuracy*. *Confusion Matrix* ditampilkan pada tabel 2 *Confusion Matrix.*

Tabel 2. *Confusion Matrix*

	Class* 1: *Positive	***Class* 2: *Negative***
Class 1: *Positive*	TP	FN
Class 2: *Negative*	FP	TN

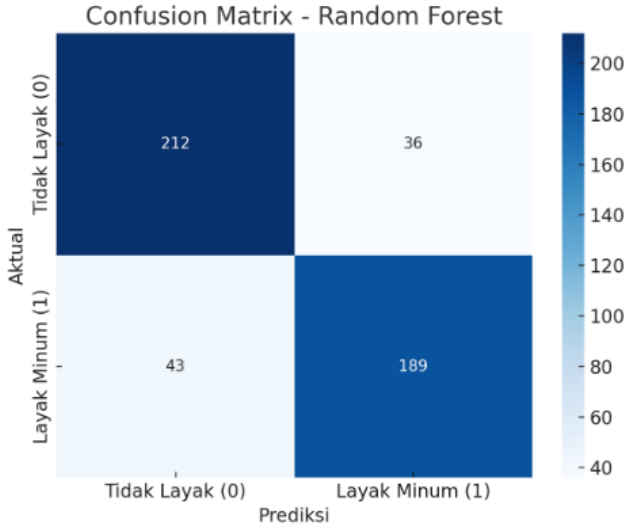
Kinerja *Confusion Matrix* dapat diukur menggunakan dengan nilai TP, FP, FN dan TN. *True Positive* merupakan data positif yang diprediksi benar. *True Negative* adalah data negatif yang diprediksi benar. *False Positive* adalah data negatif namun di prediksi sebagai data positif, *False Negative* adalah data positif namun di prediksi sebagai data negatif.

$$Accuracy = = \frac{TP + TN}{TP+FP+FN+TN} \times 100\% \quad (3)$$

3. Hasil dan Diskusi

Tahap selanjutnya evaluasi berdasarkan kedua algoritma yaitu *Random Forest* dan *Decision Tree.* kemudian akan dibuat perbandingan untuk menentukan mana yang memiliki nilai yang lebih akurat berdasarkan nilai hasil akurasi dengan *confusion matrix* dan performa yang unggul kedua algoritma tersebut.

a. Pengukuran Akurasi Algoritma *Random Forest*
Berikut *confusion matrix* yang diperoleh algoritma *Random Forest* berdasarkan pengujian menggunakan *tools python* dapat di lihat pada tabel dan gambar dibawah ini:



Gambar 4. *Confusion Matrix Random Forest*

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Tabel 3. *Confusion Matrix Random Forest*

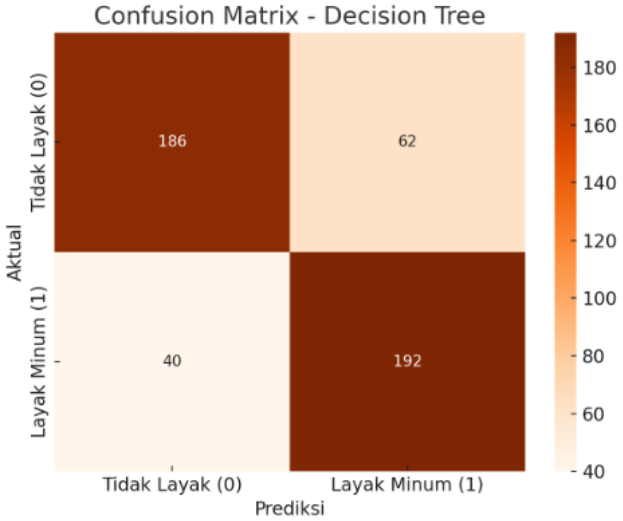
	true 0	*true 1*	*class precision*
pred. 0	212	36	85%
pred. 1	43	189	81%
Class Recall	83%	84%	

Hasil akurasi didapatkan sebesar 83.54%, dimana *class precision* untuk pred.0 (*pred.negative*) adalah 85% dan pred 1 (*pred.positive*) adalah 81%. hasil *accuracy* yang di dapatkan menggunakan persamaan 4, dimana nilai *true positive* sebanyak 189, *true negative* sebanyak 212, *false negative* sebanyak 43 dan *false positive* 36.

$$Accuracy = = \frac{189 + 212}{189+212+36+43} \times 100\% \quad (4)$$

$$Accuracy = = \frac{401}{480} \times 100\% = 83.54\% \quad (5)$$

b. Pengukuran Akurasi Algoritma *Decision Tree*
Berikut *confusion matrix* yang diperoleh algoritma *Decision Tree* berdasarkan pengujian menggunakan *tools python d*apat dilihat pada tabel dan gambar dibawah ini :



Gambar 5. *Confusion Matrix Decision Tree*

Tabel 4. *Confusion Matrix Decision Tree*

	true 0	*true 1*	*class precision*
pred. 0	186	62	75%
pred. 1	40	192	83%
Class Recall	82%	76%	

Hasil akurasi didapatkan sebesar 78.75%, dimana *class precision* untuk pred.0 (*pred.negative*) adalah 75% dan pred 1 (*pred.positive*) adalah 83%. hasil *accuracy* yang didapatkan menggunakan persamaan 4, dimana nilai *true positive* sebanyak 192, *true negative* sebanyak 186, *false negative* sebanyak 40 dan *false positive* 62.

$$Accuracy = = \frac{192 + 186}{192+186+62+40} \times 100\% \quad (6)$$

$$Accuracy = = \frac{378}{480} \times 100\% = 78.75\% \quad (7)$$

Perbandingan Perbandingan Performance Akurasi dari 2 Algoritma yaitu *Decision Tree*,dan *Random Forest*, dapat dilihat pada tabel dibawah ini:

Tabel 5. Hasil Perbandingan Akurasi

	Random Forest	*Decision Tree*
Akurasi(%)	83.54%	78.75%

Analisis hasil perbandingan akurasi *Water Quality m*enggunakan algoritma *Decision Tree* dan *Random Forest* menunjukkan bahwa *Random Forest* merupakan metode yang menghasilkan tingkat akurasi paling tinggi yaitu 83.54%, sedangkan *Decision Tree* sebesar 78.75%.

4. Kesimpulan

Tujuan dari penelitian ini untuk mengetahui keunggulan dari kedua algoritma dan hasil perbandingan tingkat akurasi prediksi dan klasifikasi kualitas air menggunakan model algoritma *machine learning* menggunakan *dataset water-potability* dari *kaggle* dengan 10 atribut dan 2 *class* yaitu *potability* dan *non-potability*. Algoritma yang digunakan yaitu *Decision Tree* dan *Random Forest*. Hasil penelitian yang telah dilakukan, menggunakan model evaluasi *confusion matrix* untuk menghitung akurasi. Akurasi merupakanyang paling popular untuk menghitung keberhasilan algoritma dalam menyelesaikan masalah, karena sebelum membuat aplikasi prediksi, sebaiknya harus mengukur kinerja algoritma yang akan digunakan, seperti dalam penelitian ini dengan membandingkan dua algoritma dan dilihat dari *Recall* dan *Precision* metode yang menghasilkan tingkat akurasi yang paling tinggi yaitu *Random Forest* sebesar 83.54%. dari hasil penelitian sebelum nya, untuk algoritma yang memiliki performa dan keunggulan untuk memantau dan pengelolaan kualitas air adalah algoritma *Random Forest* di karenakan mendapatkan akurasi yang tinggi dan algoritma Random Forest dapat digunakan untuk klasifikasi data kualitas air layak di minum.

Daftar Pustaka

[1] A. W. A. I. I. Saputra, “Implementasi Algoritma Naïve Bayes Untuk Memprediksi Kualitas Air Yang Dapat Di Konsumsi,” JATI (Jurnal Mahasiswa Teknik Informatika), no. 8, pp. 133-140, 2024.

[2] N. Maulidah[1], M. Maulidah[2], R. Supriyadi [3], H. Nalatissifa[4] dan S. Diantika[5], “Prediksi Kualitas Air Menggunakan Metode Random Forest, Decision,” Jurnal Khatulistiwa Informatika, no. 12, pp. 1-6, 2024.

[3] K. Abdi, A. Warjaya, I. Muthmainnah, dan P. H. Pahutar, “Penerapan Algoritma Random Forest dalam Prediksi Kelayakan Air Minum,” Jurnal Ilmu Komputer dan Informatika (JIKI), vol. 3, no. 2, hal. 81–88, Des. 2023.

[4] M. Meiriyama and Sudiadi, “Penerapan Algoritma Random Forest Untuk Klasifikasi Jenis Daun Herbal,” Jurnal Teknologi dan Sistem Informasi (JTSI), vol. 3, no. 1, pp. 131–138, Apr. 2022.

[5] D. Hartanti and A. I. Pradana, “Komparasi Algoritma Machine Learning dalam Identifikasi Kualitas Air,” SMARTICS Journal, vol. 9, no. 1, pp. 1–6, 2023.

[6] I. A. Putri, R. D. Saputra, and D. Andesta, “Penerapan Algoritma Random Forest untuk Prediksi Kualitas Air Bersih,” Jurnal Sistem Informasi dan Teknologi (JUSIFO), vol. 9, no. 2, pp. 129–136, Nov. 2024.

[7] G. L. P. *1, “Analisis Komparatif Algoritme Machine Learning pada,” KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, no. 2, pp. 43-55, 2022.

[8] K. Abdi, A. Warjaya, I. Muthmainnah, dan P. H. Pahutar, “Penerapan Algoritma Random Forest dalam Prediksi Kelayakan Air Minum,” Jurnal Ilmu Komputer dan Informatika (JIKI), vol. 3, no. 2, hal. 81–88, Des. 2023.
[9] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, dan U. Firdaus, “Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank,” Karimah Tauhid, vol. 3, no. 2, pp. 1860–1874, 2024.
[10] C. P. Gusti Agung Diah Sri Ari Ningsiha1, “Klasifikasi Kualitas Air Layak Minum menggunakan,” Jurnal Elektronik Ilmu Komputer Udayana, no. 1, pp. 217-226, 2024.