JURNAL KHATULISTIWA INFORMATIKA VOL. 12 NO. 1 Juni 2024
p-ISSN: 2339-1928 e-ISSN: 2579-633X Halaman : 1 - 6

PREDIKSI KUALITAS AIR MENGGUNAKAN METODE RANDOM FOREST, DECISION TREE, DAN GRADIENT BOOSTING

Nurlaelatul Maulidah[1]; Mawadatul Maulidah[2]; Riki Supriyadi[3]; Hiya Nalatissifa[4]; Sri Diantika[5]; Ahmad Fauzi[6];

Program Studi Sistem Informasi Akuntansi Kampus Kota Tegal[1,4], Program Studi Teknologi Komputer Kampus Kota Tegal[2], Program Studi Sains Data Universitas Nusa Mandiri[3], Program Studi Sistem Informasi Universitas Bina Sarana Informatika[5], Program Studi Sistem Informasi Akuntansi Kampus Kota Surakarta[6]

Universitas Bina Sarana Informatika[1][2] [4][5][6], Universitas Nusa Mandiri[3]
nurlaelatul.nlt@bsi.ac.id

INFO ARTIKEL

Diajukan :
05-06-2023

Diterima :
15-05-2024

Diterbitkan:
30-06-2024

Kata Kunci :
Kualitas Air, Prediksi, Random Forest

INTISARI

Air merupakan sumber daya alam yang sangat penting dan menjadi suatu kebutuhan pokok bagi kelangsungan mahluk hidup baik manusia, hewan dan tumbuhan, namun tidak semua air aman untuk dikonsumsi, sehingga diperlukan adanya identifikasi kualitas air yang aman untuk di konsumsi. Memperkirakan kualitas air telah menjadi salah satu tantangan signifikan yang dihadapi dunia dalam beberapa dekade terakhir. Penelitian ini menyajikan model prediksi kualitas air menggunakan tiga algoritma machine learning Decision Tree, Gradient Boosting dan Random Forest, dimana model machine learning tersebut kemudian dievaluasi secara eksperimental dengan menggunakan data water_potability dari kaggle. Ketiga algoritma ini akan dilakukan perbandingan pada proses klasifikasi data untuk mengetahui metode mana yang paling akurat, dilihat dari tingkat akurasi yang paling tinggi. Hasilnya menunjukan pada penelitian ini Random Forest menjadi model yang memiliki akurasi dengan nilai akurasi yang tinggi dan akurat sebesar 88.33%, dan untuk Decision Tree dengan nilai akurasi 80,83% dan Gradient Boosting memiliki akurasi terendah yaitu 73,33%. Sehingga pada penelitian yang dilakukan Random Forest menjadi algoritma paling akurat dan baik untuk digunakan pada dataset water_potability.

I. PENDAHULUAN

Air adalah sumber daya alam yang sangat penting dan menjadi kebutuhan dasar bagi kelangsungan hidup, baik manusia, hewan maupun tumbuhan. Air adalah senyawa yang memiliki dua bagian hidrogen (H) dan satu bagian O2 (Sahabuddin, 2015). Fungsi utama air bagi kehidupan adalah mutlak/tidak dapat digantikan karena diperlukan dalam proses fotosintesis, distribusi nutrisi dan pengatur suhu tubuh (Suyasa, 2015).

Namun seiring pertumbuhan penduduk, pertumbuhan industri, pertumbuhan ekonomi dan peningkatan standar hidup mempunyai dampak negatif terhadap sumber daya air, termasuk penurunan kualitas air itu sendiri. Selain itu, permasalahan terbesar terkait sumber daya air yaitu jumlah ketersediaan air yang sudah tidak mampu lagi memenuhi kebutuhan air yang terus meningkat dan kualitas air yang semakin turun dari tahun ke tahun. Pencemaran air merupakan salah satu masalah yang sangat serius yang perlu mendapat perhatian karena air sangat penting bagi kehidupan. adanya pencemaran air akan mengganggu kehidupan, karena setiap makhluk hidup memerlukan air yang berkualitas tinggi serta jumlah ketersediaan yang cukup. Pencemaran air adalah terjadinya perubahan dan penyimpangan sifat-sifat alami dari air yang ada di lingkungan hidup manusia (Dewata dan Danhas, 2018).

Deteksi dini yang dapat dijadikan masukan melalui teknik *data mining* diperlukan untuk mengetahui kualitas air yang layak dikonsumsi masyarakat luas, dalam penelitian ini teknik klasifikasi yang digunakan bertugas untuk melakukan prediksi suatu kategori label serta membedakan objek satu dengan objek lainnya (Said, Matondang dan Irmanda, 2022). Data Mining adalah disiplin ilmu yang digunakan untuk memecahkan masalah pengambilan data dari *database* yang besar dengan menggabungkan Teknik dari statistik, pembelajaran mesin, visualisasi data, pengenlan pola, dan database (Werdiningsih, Nuqoba & Muhammadun, 2020) .

Penelitian ini menggunakan *dataset Water Potability* dengan tipe *file* csv, berisi data kualitas air dengan *value* yang berbeda-beda pada 10 *attributte* yang dimiliki, *attributte potability*

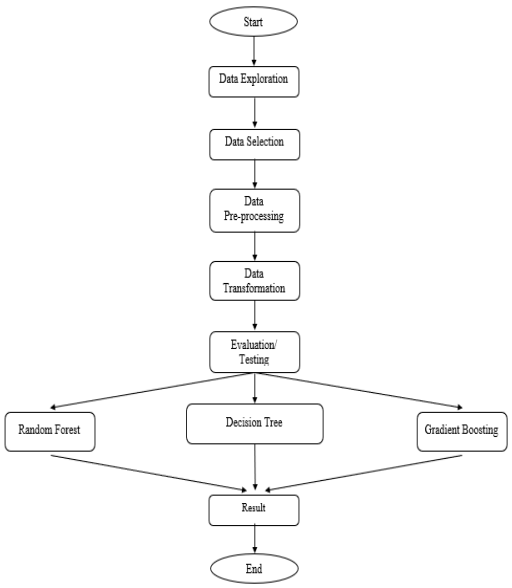
merupakan *attributte* yang diberi label sebagai hasil *conclusion* apakah air tersebut aman atau tidak untuk digunakan berdasarkan hasil akurasi dari masing-masing *value attributte* yang ada pada data tersebut.

Berdasarkan permasalahan yang ada yaitu melakukan perbandingan ketiga metode : *Random Forest*, *Decision Tree* dan *Gradient Boosting* maka dilakukan penelitian dengan judul "Prediksi Kualitas Air Menggunakan Metode *Random Forest*, *Decision Tree*, Dan *Gradient Boosting"* menggunakan *tools Google Colab* untuk mengetahui nilai akurasi yang paling besar dari ketiga metode yang akan diimplementasikan kedalam klasifikasi data. Tujuan penelitian ini untuk membandingkan diantara ketiga metode yang paling baik digunakan dalam klasifikasi kualitas air dengan hasil akurasi yang paling maksimal.

II. BAHAN DAN METODE

Penelitian ini merupakan penelitian eksperimen. Pengumpulan data dalam penelitian ini meliputi studi literatur berupa buku, jurnal, dan karya ilmiah yang relevan mengenai prediksi kualitas air. Dataset yang digunakan adalah *Water Potability Dataset* dari Kaggle, dimana terdapat 3276 record, 10 atribut dan 2 target kelas yang kemudian akan diolah menggunakan *tools Google Colabs* dan data akan dibagi untuk *data training* dan *data testing* dengan *Python Programming* menggunakan algoritma *random forest*, *decision tree*, dan *gradient boosting*. Pada dataset terdapat 10 atribut antara lain : ph, *Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, Potability*.

Berikut langkah-langkah untuk mendapatkan nilai akurasi pada penelitian ini adalah sebagai berikut:



Gambar 1. Tahapan Penelitian

a. Data Exploration

Pada penelitian ini data yang digunakan yaitu data sekunder. Data sekunder merupakan sumber data yang didapatkan peneliti dengan media perantara atau tidak secara langsung. Data sekunder pada penelitian ini menggunakan *Dataset Water Potability* dari *Kaggle.*

b. Data Selection

Data Selection merupakan proses menganalisis data yang sesuai dari *database*, karena tidak semua data dibutuhkan dalam proses data mining.

c. Data Pre-processing

Data preprocessing ini merupakan langkah awal dalam melakukan prediksi kualitas air. Tahapan *preprocessing* data didalam penelitian ini terdiri dari Pemeriksaan data kosong, menghapus duplikasi data serta membuang data yang tidak konsisten dan *noise* pada setiap atribut *dataset*.

d. Data Transformation

Transformasi data merupakan proses perubahan atau penggabungan data ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa algoritma *data mining* membutuhkan format data yang khusus sebelum diaplikasikan. Data yang *non-numeric* akan dilakukan proses perubahan/inisiasi ke dalam bentuk *numeric*. Namun jika data yang dimiliki sudah dalam bentuk *numeric* maka tidak diperlukan inisiasi.

e. Evaluation/Testing

Dalam penelitian ini, pemodelan dataset dilakukan dengan menggunakan algoritma *random forest*, *decision tree*, dan *gradient boosting* pada *Dataset Water Potability* dari Kaggle Pembahasan lebih lanjut akan dijelaskan pada bagian pembahasan dan kesimpulan. Pengujian model akan menggunakan *train test split*, evaluasi yang akan dihasilkan yaitu nilai akurasi dari ketiga metode tersebut, setelah itu ketiga metode tersebut dibandingkan untuk mendapatkan algoritma prediksi kualitas air yang akurat.

Ada banyak matrik yang dapat digunakan untuk mengukur kinerja *classifier* atau *predictor*, bidang yang berbeda memiliki preferensi yang berbeda untuk metrik tertentu karena tujuan yang berbeda (Chicco and Jurman, 2020).

Confusion matriks adalah hasil prediksi dari masalah klasifikasi (Siregar, 2020).

Tabel 1. *Confusion Matrix*

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

Sumber : (Siregar, 2020)

Dimana:
*Class*1 = *Positive*; *Class*2 = *Negative*;
TP = *True Positive*. TN = *True Negative*
FP = *False Positive*, FN = *False Negative*.
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 negative yang diprediksi benar.
- *False Positive* adalah data negatif namun diprediksi sebagai data positif.
- *False Negative* adalah data positif namun diprediksi sebagai data *negative*.

Berikut formula untuk menghitung akurasi:

$$Akurasi\ (\%) = \frac{TP + TN}{TP + TN + FP + FN}\ x\ 100 \qquad (1)$$

III. HASIL DAN PEMBAHASAN

Dataset Water Quality diperoleh dari situs kaggle (www.kaggle.com/adityakadiwal/water-potability) (Kadiwal, 2021). Pada penelitian ini menggunakan dataset "*water potability*" atau kualitas air dengan tipe file csv untuk proses klasifikasi dalam membandingkan hasil akurasi dari ketiga metode yang digunakaan yaitu *Decision Tree*, *Gradient Boosting* dan *Random Forest*.

Dataset water_potability memiliki jumlah data 3276 baris dengan jumlah atribut sepuluh yang gunakan dalam penelitian ini, yaitu:

1. ph: pH 1 air (0 sampai 14)
2. *Hardness* (Kekerasan): kapasitas air untuk mengendapkan sabun dalam mg/L
3. *Solids* (Padatan): Total padatan terlarut dalam satuan ppm
4. *Chloramines* (Kloramin): Jumlah Kloramin dalam satuan ppm
5. *Sulfate* (Sulfat): Jumlah Sulfat yang terlarut dalam satuan ppm
6. *Conductivity* (Konduktivitas): Konduktivitas listrik air dalam satuan µS/cm
7. *Organic Carbon* (Karbon Organik): Jumlah karbon organik dalam satuan ppm
8. *Trihalomethanes* (Trihalomethanes): Jumlah *Trihalomethanes* dalam satuan µg/L
9. *Turbidity* (Kekeruhan): Ukuran properti pemancar cahaya air dalam satuan NTU
10. *Potability* (sifat dapat diminum): menunjukkan apakah air aman untuk dikonsumsi manusia, dapat diminum 1 dan tidak dapat diminum 0. Atribut ph sampai dengan kekeruhan merupakan tipe data *real* dan atribut potabilitas merupakan tipe data integer.

Data Eksploration

Data Eksplorasi dilakukan untuk menganalisis data awal. Dimana ditampilkan secara visual penambangan untuk memahami apa yang ada dalam kumpulan data dan karakteristik data. Karakteristik ini mungkin termasuk ukuran atau kuantitas data, kelengkapan data, keakuratan data, kemungkinan hubungan antara variabel, file atau tabel data. Data dianalisis menggunakan python dengan google colab.

Pre-processing Data

Data yang diperoleh pada penelitian ini perlu dilakukan proses *pre-processing*. Proses labeling data dalam penelitian ini yaitu potability, merupakan attribute untuk mengetahui apakah air aman dikonsumsi (1) atau tidak (0).

Beberapa *preprocessing* yang digunakan seperti resampling dilakukan pada dataset untuk menghilangkan masalah ketidakseimbangan data. Kemudian menghilangkan *missing value* (nilai null), yang membuat perhitungan menjadi lebih mudah.

Selain itu, juga mereduksi dimensi dataset yang dipengaruhi oleh variabel multikorelasi, atribut-atribut data serta transformasi data kemudian dilakukan proses pembersihan dan regulasi data dan lain sebagainya.

Pada penelitian ini setelah dilakukan proses *pre-processing* data maka data yang sebelumnya 3276 data mejadi 2400 data.

Data-Split

Pada penelitian ini dataset dibagi menjadi dua set yaitu *data training* 0.9 (90%) dan data testing 0.1 (10%), menghasilkan sebesar 2160 untuk *data training* dan 240 data untuk *data testing*.

Evaluation

Pada tahap ini dilakukan evaluasi berdasarkan ketiga algoritma yaitu *Gradient Boosting, Decision Tree* dan *Random Forest*. Kemudian dibuat perbandingan untuk menentukan mana yang memiliki nilai yang lebih akurat berdasarkan nilai dari hasil akurasi dengan *confusion matrix* dan juga *tools* data mining yang digunakan yaitu google colab dengan Python.

1. Pengukuran Akurasi Algoritma Decision Tree

Berikut *confusion matrix* yang diperoleh algoritma *Decision Tree* berdasarkan pengujian menggunakan *tools Python* dapat dilihat pada Tabel dibawah ini:

Tabel 2. *Confusion Matrix* Algoritma *Decision Tree*
Akurasi : 81%

	True 0	True 1	Class precission
Pred. 0	83	26	81%
Pred. 1	20	111	81%
Class recall	76%	85%	

Sumber: penelitian (2022)

Hasil akurasi sebesar 81%, dengan *class precision* untuk pred. 0 (pred. negative) dan pred 1 (pred.positive). Hasil *accuracy* didapatkan menggunakan persamaan 1, dimana nilai *true positive* sebanyak 83, *true negative* sebanyak 111, *false negative* sebanyak 26, dan *false positive* 20. Hasil akurasi dapat dibuktikan dengan:

$$Akurasi\ (\%) = \frac{83 + 111}{83 + 111 + 20 + 26}\ x\ 100$$

$$Akurasi\ (\%) = \frac{194}{240}\ x\ 100 = 80{,}83\ \%$$

Performance Vektor Algoritma *Decision Tree* dapat dilihat pada tabel berikut:

Tabel 3. Hasil *Performance* Vektor *Decision Tree*

Performance Vector
Accurancy 80,83%
Confusion Matrix

True	0	1
0	83	26
1	20	111

Sumber: penelitian (2022)

Performance Vector adalah deskripsi dari tabel hasil analisis yang diperoleh dalam penelitian yang dilakukan.

- Nilai *True Positive* sebanyak 83, yaitu nilai data positif yang artinya air tersebut aman untuk diminum dan diprediksi memiliki nilai yang benar.
- Nilai *False Positive* sebanyak 20, dimana data negatif (air tidak dapat diminum) namun diprediksi sebagai data positif.
- Nilai *False Negative* sebanyak 26, data positif namun diprediksi sebagai data negatif.
- Nilai *True Negative* sebanyak 111, merupakan data negatif yang diprediksi benar.

2. Pengukuran Akurasi Algoritma *Gradient Boosting*

Berikut *confusion matrix* yang diperoleh algoritma *Gradient Boosting* berdasarkan pengujian menggunakan *tools* Python dapat dilihat pada Tabel dibawah ini:

Tabel 4. *Confusion Matrix Gradient Boosting*
Akurasi : 73%

	True 0	True 1	Class precission
Pred. 0	75	34	71%
Pred. 1	30	101	75%
Class recall	69%	77%	

Sumber: penelitian (2022)

Hasil akurasi sebesar 73%, dengan class precision untuk pred. nol (pred. negative) adalah 71% dan pred satu (pred.positive) adalah 75%. Hasil accuracy didapatkan menggunakan persamaan 1, dimana nilai *true positive* sebanyak 75, *true negative* sebanyak 101, *false negative* sebanyak 34, dan *false positive* 30. Hasil akurasi dapat dibuktikan dengan:

$$Akurasi\ (\%) = \frac{75 + 101}{75 + 101 + 30 + 34}\ x\ 100$$

$$Akurasi\ (\%) = \frac{176}{240}\ x\ 100 = 73{,}33\%$$

Performance Vektor Algoritma *Gradient Boosting* dapat dilihat pada table berikut:

Tabel 5. Hasil *Performance* Vektor *Gradient Boosting*

Performance Vector
Accurancy 73,33%
Confusion Matrix

True	0	1
0	75	34
1	30	101

Sumber: penelitian (2022)

Performance Vector merupakan bentuk deskripsi dari tabel hasil analisis yang diperoleh dalam penelitian yang dilakukan.

- Nilai *True Positive* sebanyak 75, merupakan nilai data positif yang artinya air aman untuk diminum dan diprediksi memiliki nilai yang benar.
- Nilai *False Positive* sebanyak 30, dimana data negatif (air tidak dapat diminum) namun diprediksi sebagai data positif.
- Nilai *False Negative* sebanyak 34, data positif namun diprediksi sebagai data negatif.
- Nilai *True Negative* sebanyak 101, merupakan data negatif yang diprediksi benar.

3. Pengukuran Akurasi Algoritma *Random Forest*

Berikut *confusion matrix* yang diperoleh algoritma *Random Forest* berdasarkan pengujian menggunakan *tools* Python dapat dilihat pada Tabel dibawah ini:

Tabel 6. *Confusion Matrix Random Forest*
Akurasi : 88%

	True 0	True 1	*Class precission*
Pred. 0	96	13	86%

Pred. 1	15	116	90%
Class recall	88%	89%	

Sumber: penelitian (2022)

Hasil akurasi sebesar 88%, dengan *class precision* untuk pred. nol (pred. *negative*) adalah 86% dan pred satu (pred.*positive*) adalah 90%. Hasil *accuracy* didapatkan menggunakan persamaan 1, dimana nilai *true positive* sebanyak 96, *true negative* sebanyak 116, false negative sebanyak 13, dan *false positive* 15. Hasil akurasi dapat dibuktikan dengan:

$$Akurasi\ (\%) = \frac{96 + 116}{96 + 116 + 15 + 13}\ x\ 100$$

$$Akurasi\ (\%) = \frac{212}{240}\ x\ 100 = 88{,}33\ \%$$

Performance Vektor Algoritma Random Forest dapat dilihat pada tabel dibawah ini:

Tabel 7. Hasil Performance Vektor Random Forest

Performance Vector Accurancy 88,33% Confusion Matrix		
True	0	1
0	96	13
1	15	116

Sumber: penelitian (2022)

Performance Vector merupakan bentuk deskripsi dari tabel hasil analisis yang diperoleh dalam penelitian yang dilakukan.

- Nilai *True Positive* sebanyak 96, merupakan nilai data positif yang artinya air aman untuk diminum dan diprediksi memiliki nilai yang benar.
- Nilai *False Positive* sebanyak 15, dimana data negatif (air tidak dapat diminum) namun diprediksi sebagai data positif.
- Nilai *False Negative* sebanyak 13, data positif namun diprediksi sebagai data negatif.
- Nilai *True Negative* sebanyak 116, merupakan data negatif yang diprediksi benar.

Perbandingan *Performance* Akurasi dari 3 Algoritma yaitu *Decision Tree*, *Gradient Boosting* dan *Random Forest*, dapat dilihat pada tabel dibawah ini:

Tabel 8. Hasil Perbandingan Akurasi

	Decision Tree	***Gradient Boosting***	***Random Forest***
Akurasi(%)	80,83	73,33	88,33

Analisis hasil perbandingan akurasi *Water Quality* menggunakan algoritma *machine learning Decision Tree, Gradient Boosting* dan *Random Forest* menunjukkan bahwa *Random Forest* merupakan metode yang menghasilkan tingkat akurasi paling tinggi yaitu 88,33%, sedangkan *Decision Tree* sebesar 80.83% dan *Gradient Boosting* sebesar 73.33%.

IV. KESIMPULAN

Tujuan dari penelitian ini untuk mengetahui 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*. *Machine Learning* yang digunakan yaitu *Decision Tree*, *Gradient Boosting* 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 tiga algoritma dan dilihat dari *Recall* dan *Precision* metode yang menghasilkan tingkat akurasi yang paling tinggi yaitu *Random Forest* sebesar 88.33%. Metode klasifikasi *Decision Tree* dan *Gradient Boosting* pada penelitian ini cukup baik digunakan karena menghasilkan tingkat akurasi diatas 70%, namun untuk mendapatkan hasil akurasi yang lebih maksimal untuk kedepannya dapat menggunakan metode yang lain.

V. REFERENSI

D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

Dewata, I. & Danhas, Y. H. (2018). *Pencemaran Lingkungan*. Depok: Rajawali Pres.

Kadiwal, A., 2021. "Water Quality" (Online), (https://www.kaggle.com/adityakadiwal/water-potability, diakses 7 September 2022).

Sahabuddin, E. S. (2015). *Filosofi 'Cemaran' Air*. Kupang: PTK Press.

Siregar, A. (2020). Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning. PETIR, 13(2), 138 - 147. https://doi.org/10.33322/petir.v13i2.998

Suyasa, W. B. (2015). *Pencemaran Air & Pengolahan Air Limbah*. Denpasar: Udayana University Press.

Said, H., Matondang, N., & Irmanda, N. (2022).

Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi. *Techno.COM, Vol. 21, No. 2, Mei 2022: 256-267*

Werdiningsih, I., Nuqoba, B., & Muhammadun (2020). *Data Mining Menggunakan Android, Weka, Dan SPSS*. Surabaya: Penerbit Airlangga University Press.