JTRISTE, Vol.11, No.1, Maret 2024, pp. 53~59
ISSN: **2355-3677 | e-ISSN: 2460-8548**

Perbandingan Algoritma Random Forest dan Artificial Neural Network untuk Dataset Water Potability

M. Rizky Fajar Mali[1], Basilius Arilla Dimas N.[2], Ajitama Jaya[3], Iwan Binanto[4*]
[1,2,3,4] Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma
e-mail: [1]rizkyfadjarmali@gmail.com, [2]basiliusarilla@gmail.com, [3]anjitama38@gmail.com, [4]iwan@usd.ac.id

Abstrak

Air merupakan elemen penting yang menjadi kebutuhan mendasar bagi kelangsungan hidup semua makhluk hidup. Saat ini, kesadaran masyarakat tentang pentingnya air berkualitas dan bermutu semakin meningkat. Hal ini disebabkan oleh pemahaman yang lebih luas tentang dampak kesehatan dari air yang tidak bersih. Air yang bersih dan aman untuk digunakan tidak hanya berdampak positif pada kesehatan kita, tetapi juga pada berbagai aspek kehidupan sehari-hari. Oleh karena itu, penelitian tentang kualitas dan kelayakan air untuk konsumsi sangat penting. Tujuan penelitian ini adalah untuk menentukan metode terbaik untuk membandingkan kualitas air dalam kelayakan konsumsi. Penelitian ini membandingkan dua metode machine learning, yaitu Random Forest dan Artificial Neural Network (ANN) berdasarkan atribut dalam kelayakan air minum, yakni: PH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity dan potability. Hasil penelitian menunjukkan algoritma Random Forest memiliki tingkat akurasi sebesar 67.823%, sedangkan algoritma Artificial Neural Network (ANN) mencapai akurasi 61.014%. Dari hasil ini, dapat disimpulkan bahwa algoritma Random Forest memiliki akurasi lebih tinggi dibandingkan dengan Artificial Neural Network (ANN).

Kata kunci: *Perbandingan Algoritma, Random Forest, Artificial Neural Network, Water Potability.*

Abstract

Water is an important element that is a basic need for the survival of all living things. Currently, public awareness about the importance of good quality water is increasing. This is due to a wider understanding of the health impacts of unclean water. Water that is clean and safe to use not only has a positive impact on our health, but also on various aspects of daily life. Therefore, research on the quality and suitability of water for consumption is very important. The aim of this research is to determine the best method for comparing water quality for suitability for consumption. This research compares two machine learning methods, namely Random Forest and Artificial Neural Network (ANN) based on attributes for the suitability of drinking water, namely: PH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity and potability. The research results show that the Random Forest algorithm has an accuracy rate of 67.823%, while the Artificial Neural Network (ANN) algorithm achieves an accuracy of 61.014%. From these results, it can be concluded that the Random Forest algorithm has higher accuracy compared to Artificial Neural Network (ANN).

Keywords*: Algorithms Comparison,Random Forest, Artificial Neural Network, Water Potability*

1. Pendahuluan

Akses terhadap air minum yang aman sangat penting bagi kesehatan, merupakan hak semua orang dan merupakan komponen kebijakan yang efektif untuk kesehatan. Hal ini penting

* Corresponding author : Iwan Binanto (iwan@usd.ac.id)

Diterima *: April, 2024* ***Disetujui*** *: April, 2024* ***Dipublikasikan*** *: April, 2024*

sebagai isu kesehatan dan pembangunan di tingkat nasional, regional dan lokal. Di beberapa wilayah, investasi pada penyediaan air dan sanitasi terbukti dapat menghasilkan manfaat ekonomi, karena pengurangan dampak buruk terhadap kesehatan dan biaya pelayanan kesehatan lebih besar daripada biaya yang dikeluarkan untuk melakukan intervensi [1].

Peningkatan polusi dan perubahan lingkungan dapat mempengaruhi parameter kualitas air seperti pH, kadar logam, dan bahan organik. Beberapa penelitian sudah dilakukan untuk menentukan air layak dikonsumsi dengan Teknik klasifikasi machine learning [2], [3].
Penelitian ini bertujuan untuk mengevaluasi kualitas air sebagai layak atau tidaknya untuk dikonsumsi menggunakan pendekatan algoritma klasifikasi, yaitu Random Forest dan Artificial Neural Network (ANN).

2. Tinjauan Pustaka

Klasifikasi adalah sebuah proses penting yang melibatkan pencarian pola-pola tertentu dengan tujuan utama untuk memperkirakan kelas atau kategori dari objek yang statusnya belum diketahui atau belum ditentukan. Proses ini memegang peranan penting dalam berbagai bidang, termasuk ilmu komputer dan data mining [4]. Secara lebih rinci, klasifikasi dapat diartikan sebagai metode atau pendekatan sistematis untuk mengidentifikasi dan membuktikan bahwa sebuah objek data tertentu merupakan bagian dari satu jenis atau kategori yang telah dideskripsikan atau ditentukan sebelumnya [5].

Pengelompokan dalam klasifikasi dilakukan berdasarkan ciri-ciri atau karakteristik tertentu yang dimiliki oleh suatu objek. Proses ini bisa dilakukan oleh manusia secara manual, atau dengan bantuan teknologi dan alat-alat canggih yang ada saat ini [6]. Salah satu teknik klasifikasi yang populer dan sering digunakan adalah teknik yang menggunakan algoritma Naive Bayes. Algoritma ini memiliki keunggulan dalam hal efisiensi dan efektivitas, menjadikannya pilihan favorit dalam berbagai aplikasi klasifikasi.

Artificial Neural Network (ANN) adalah tipe model untuk pembelajaran mesin yang semakin berkembang kompetitif terhadap model statistik dan regresi konvensional dalam hal kegunaan. Berdasarkan faktor analisis data termasuk pengolahannya kecepatan, latensi, toleransi kesalahan, kinerja, volume, akurasi konvergensi, dan skalabilitas, penerapan ANN secara penuh dapat dinilai. Karena potensi besar dari ANN yang merupakan pemrosesan berkecepatan tinggi yang ditawarkan dalam implementasi paralel yang masif. ANN dapat dikembangkan dan digunakan dalam pemrosesan bahasa alami, pengenalan gambar, dll. Berkat fitur luar biasa dari kemampuan beradaptasi, pembelajaran mandiri, toleransi kesalahan, dan mampu melakukan pemodelan nonlinier tanpa pengetahuan sebelumnya tentang hubungan antara variabel masukan dan keluaran, menjadikannya alat peramalan yang lebih umum dan fleksibel [7]. ANN telah digunakan dalam perkiraan fungsi di seluruh dunia dalam paradigma numerik. Faktor analisis data tersebut menjelaskan lebih jelas efektivitas, efisiensi, dan

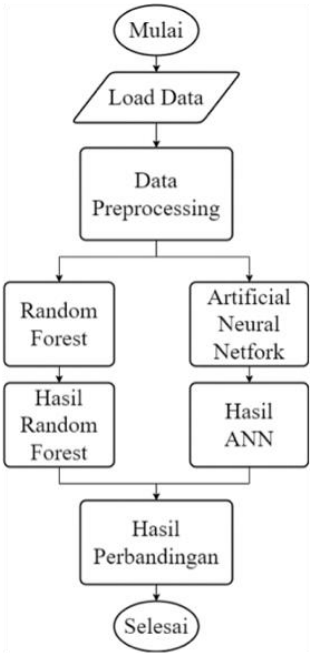
keberhasilan ANN dalam menawarkan kemampuan tinggi untuk menyelesaikan permasalahan yang tidak rumit dan kompleks dalam berbagai bidang kehidupan [8], [9].

Random Forest adalah teknik penambangan data prediktif dan juga merupakan metode machine learning ensemble. Konsep utama dari metode ensemble adalah sekelompok 'weak learners' (tree) bergabung bersama untuk membentuk 'strong learners'(random forest). Random Forest adalah kombinasi dari beberapa decision tree yang digabungkan untuk memperoleh prediksi yang akurat [10]. Dari beberapa penelitian yang menggunakan dataset Water Potability untuk membandingkan beberapa algoritma, hasil akurasi yang didapatkan sangat bervariasi. Hasil penelitian ANN menggunakan beberapa metode imputation yaitu Menghapus Missing Data, Median atau Mean Imputation, Arbitrary Value Imputation, KNN Imputation menghasilkan nilai akurasi yang tinggi menggunakan algoritma KNN Imputation dan Mean Median Imputation [2].

Perbandingan algoritma KNN dan SVM dalam klasifikasi menghasilkan akurasi sebesar 69,764% untuk algoritma SVM [4]. untuk penelitian menggunakan algoritma Random Forest, LightGBM, dan Decision Tree tanpa Hyperparameter tuning menghasilkan akurasi 0.824695 untuk Random Forest, 0.807927 untuk LightGBM, dan 0.746951 untuk Decision Tree. setelah dilakukan Hyperparameter tuning nilai akurasi berubah menjadi 0.792683 untuk Random Forest, 0.795732 untuk LightGBM, dan 0.766768 untuk Decision Tree. dari penelitian tersebut dapat disimpulkan bahwa algoritma yang yang menghasilkan nilai akurasi tertinggi adalah algoritma Random Forest tanpa Hyperparameter dengan nilai akurasi 0.824695 [11]. Dari penelitian perbandingan berbagai macam algoritma didapatkan hasil bahwa Random Forest menghasilkan nilai akurasi tertinggi dibandingkan dengan Artificial Neural Network (ANN) dan Naive Bayes dengan nilai akurasi 72.45% [12]. Penelitian Lidia Savitri menunjukkan bahwa algoritma Random Forest Classifier memiliki akurasi lebih tinggi daripada algoritma lain yang digunakan. Sistem identifikasi kualitas air minum memiliki akurasi 76.52% dengan algoritma Random Forest Classifier [13]. Hasil penelitian Aldi Tangkelayuk menunjukkan bahwa metode K-nearest Neighbors memiliki tingkat akurasi tertinggi sebesar 86,88%, dibandingkan dengan Decision Tree sebesar 80,84% dan Naïve Bayes sebesar 63,60%. Oleh karena itu, metode K-nearest Neighbors merupakan metode terbaik untuk klasifikasi data [14]. Berdasarkan hasil studi literatur, G. L. Pritalia menyimpulkan bahwa metode K-Nearest Neighbor dan Naïve Bayes memiliki tingkat akurasi yang cukup tinggi. Dalam pengujian, diperoleh nilai akurasi untuk metode KNN sebesar 82.42% dan nilai akurasi untuk metode Naïve Bayes sebesar 70.32% [15].

3. Metode Penelitian

Penelitian ini dilakukan dengan menggunakan dataset yang diambil dari Kaggle dengan nama Water Potability yaitu memprediksi apakah suatu air dapat layak untuk diminum atau tidak. Algoritma yang digunakan adalah Random Forest dan Artificial Neural Network, yang akan dibandingkan untuk melihat efektifitas algoritma-algoritma tersebut. Penelitian ini menggunakan suatu metode, metode tersebut dapat dilihat pada Gambar 1.



Gambar 1. Metode Penelitian

Dataset ini memiliki sembilan fitur dan satu target, yaitu:

1) ph, merupakan fitur yang mengevaluasi asam-basa air
2) *Hardness*, kapasitas air untuk mengendapkan sabun dalam mg/L.
3) *Solids*, total padatan terlarut dalam ppm.
4) *Chloramines*, jumlah Kloramin dalam ppm.
5) *Sulfate*, jumlah Sulfat yang dilarutkan dalam mg/L.
6) *Conductivity*, konduktivitas listrik air dalam µS/cm.
7) *Organic_carbon*, jumlah karbon organik dalam ppm.
8) *Trihalomethanes*, jumlah *Trihalomethanes* dalam µg/L.
9) *Turbidity*, ukuran sifat memancarkan cahaya air di NTU.
10) *Potability*, menunjukkan apakah air aman untuk dikonsumsi manusia atau tidak. Dapat diminum : 1 dan Tidak dapat diminum : 0. (Merupakan target)

Sedangkan parameter-parameter dari algoritma klasifikasi yang digunakan seperti terlihat pada Tabel 1.

Tabel 1: Parameter yang digunakan pada masing-masing Algoritma Klasifikasi

Klasifikasi	Parameter
Random Forest (RF)	criterion='entropy', max_features=None, n_estimators=150, random_state=0
Artificial Neural Network (ANN)	model = sequential Parameter layer menggunakan 8 layer model.add(Dense(units=10,activation='relu')) model.add(Dense(units=8,activation='relu')) model.add(Dense(units=8,activation='relu')) model.add(Dense(units=6,activation='relu')) model.add(Dense(units=6,activation='tanh')) model.add(Dense(units=5,activation='relu')) model.add(Dense(units=1,activation='tanh')) model.compile(loss='binary_crossentropy', optimizer='adam') model.fit(x=X_train,y=y_train,epochs=500,validation_data=(X_test, y_test), verbose=1)

Hasil dari dua algoritma akan dibandingkan untuk evaluasi akurasi, precision, recall, F1-Score dan waktu pada data. Proses preprocessing dilakukan sebelum klasifikasi, termasuk pengecekan missing value dan normalisasi data. Setelah preprocessing, melakukan split data test sebesar 30% dan data train sebesar 70%, kemudian melakukan modeling untuk Random Forest dan ANN. Spesifikasi laptop yang digunakan pada penelitian ini tertera pada Tabel 2.

Tabel 2: Spesifikasi Hardware yang digunakan

Bagian	Keterangan
Processor (CPU)	Intel Core i5-8250U
Ram	8 Gb
Storage	512 Gb
Graphic (GPU)	NVIDIA GeForce MX130

4. Hasil dan Pembahasan

Penelitian ini menggunakan dataset Water Potability yang diperoleh dari situs Kaggle. Dataset yang diperoleh dilakukan proses preprocessing yang dimana dilakukan pengecekan missing value, dan normalisasi. Dataset memiliki missing value pada 3 feature yaitu ph, Sulfate, dan Trihalomethanes, dan dilakukan impute pada missing value, data yang memiliki missing value diinputkan dengan median dari setiap feature. Untuk proses normalisasi menggunakan MinMaxScaller, normalisasi ini dilakukan pada saat sesudah melakukan proses split train dan test data. Untuk data train digunakan sebesar 70% dan data test sebesar 30%. Pengujian dilakukan dengan algoritma Random Forest dan Artificial Neural Network (ANN) dengan parameter algoritma yang tertera pada Tabel 1. Perbandingan kedua algoritma ini terlihat pada Tabel 3.

Tabel 3: Hasil Perbandingan

	Random Forest	ANN
Precision	0.678	0.610
Recall	0.678	0.610
F1 Score	0.678	0.610
Accuracy	67.8237650%	61.0146862%
Runtime	7.2 detik	47.4 detik

Random Forest menghasilkan nilai precision, recall, F1-Score adalah 0.678 dan accuracy 67.8237650% dengan runtime 7,2 detik. Kemudian pada algoritma Artificial Neural Network (ANN) menghasilkan nilai precision, recall, F1-Score adalah 0.610 dan accuracy adalah 61.0146862% dengan runtime 47,4 detik.

5. Kesimpulan

Dari penjabaran data yang telah dilakukan, terlihat bahwa algoritma yang digunakan dapat digunakan untuk memprediksi apakah air layak dikonsumsi dengan menggunakan dataset Water Potability. Hasil yang diperoleh dari setiap algoritma sangat beragam dan didokumentasikan dalam bentuk Precision, Recall, F1-Score, accuracy, dan Runtime. Jika melihat data perbandingan hasil, dapat disimpulkan bahwa algoritma Random Forest menjadi algoritma terbaik untuk melakukan prediksi dengan dataset ini.

Namun, algoritma Random Forest belum tentu menjadi algoritma terbaik untuk melakukan prediksi dengan dataset ini. Kami melakukan pengujian dataset menggunakan algoritma Random Forest dan Artificial Neural Network (ANN). Setelah penelitian, kami mendapatkan nilai akurasi yang cukup rendah, yaitu 67.8237650% menggunakan algoritma Random Forest dan 61.0146862% menggunakan algoritma Artificial Neural Network (ANN). Kami merasa perlu untuk dilakukan penelitian ulang dengan menggunakan parameter-pareameter yang berbeda untuk mendapatkan nilai akurasi yang lebih tinggi.

Maka dalam makalah ini, disimpulkan bahwa algoritma Random Forest adalah algoritma paling baik digunakan untuk melakukan prediksi apakah suatu air layak diminum atau tidak, jika dibandingkan dengan menggunakan algoritma Artificial Neural Network (ANN).

Daftar Pustaka

[1] A. Kadiwal, “Water Quality Drinking water potability.” [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability.

[2] R. Yang, “Analyses of Approaches to Deal with Missing Data in Water Quality Data Set,” 2022.

[3] S. Ulum, R. Fahmi Alifa, P. Rizkika, and C. Rozikin, “Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum,” 2023.

[4] M. Yoshe and W. Hadikurniawati, “Implementasi Metode Naive Bayes Classifier Untuk Klasifikasi Status Gizi Stunting Pada Balita,” vol. 2019, 2021.

[5] F. Alghifari and D.Juardi, “Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes,” 2021.

[6] M. M. Mutoffar *et al.*, “Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest,” vol. 04, 2022.

[7] J. C. Tellez Gaytan *et al.*, “AI-Based Prediction of Capital Structure: Performance Comparison of ANN SVM and LR Models,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/8334927.

[8] A. Tuan Hoang *et al.*, “A review on application of artificial neural network (ANN) for performance and emission characteristics of diesel engine fueled with biodiesel-based fuels,” *Sustainable Energy Technologies and Assessments*, vol. 47, Oct. 2021, doi: 10.1016/j.seta.2021.101416.

[9] R. Dastres and M. Soori, “Artificial Neural Network Systems,” 2021. [Online]. Available: https://www.researchgate.net/publication/350486076

[10] Soumi Ghosh and Chandan Banerjee, “A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment,” 2020.

[11] E. Arslan, “Water Potability Prediction.” [Online]. Available: https://www.kaggle.com/code/emrearslan123/water-potability-prediction/notebook

[12] A. Anwar, “Water Quality | LuciferML | 76% | Deployment.” [Online]. Available: https://www.kaggle.com/code/d4rklucif3r/water-quality-luciferml-76-deployment

[13] L. Savitri and R. Nursalim, “Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning.” [Online]. Available: https://ejournal.unib.ac.id/diophantine,

[14] A. Tangkelayuk and E. Mailoa, “Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree,” vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[15] G. L. Pritalia, “Analisis Komparatif Algoritme Machine Learning pada Klasifikasi Kualitas Air Layak Minum,” 2022.