PENERAPAN ALGORITMA DATA MINING UNTUK KLASIFIKASI KUALITAS AIR

Fauzi Yusa Rahman [1)], Indu Indah Purnomo[2)], Nadya Hijriana[3)]

[1]Fakultas Teknologi Informasi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari
email : fauziyusarahman@gmail.com
[2]Fakultas Teknologi Informasi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari
email : indumbc@gmail.com
[3]Fakultas Teknologi Informasi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari
email : nadyahijriana@gmail.com

Abstrak

Air merupakan sumber kehidupan terutama bagi kehidupan manusia, maka dari itu kuualitas air tersebut wajib dijaga demi keberlangsungan kehidupan manusia dan alam sekitarnya. Adanya pembangunan yang semakin pesat mendorong banyaknya penggunaan lahan di sepanjang aliran sungai. Hal ini bisa dilihat, terutama sungai-sungai di perkotaan, yang berubah fungsi menjadi permukiman dan kegiatan industri. Perubahan fungsi ini menimbulkan kekhawatiran akan menurunnya kualitas air yang mengalir sepanjang sungai tersebut. Adapun tujuan dari penelitian ini adalah menerapkan metode data mining untuk mengklasifikasi kualitas air berdasarkan parameter kualitas air meliputi mikrobiologi, kimia anorganik, dan parameter kimia. Beberapa algoritma yang diuji performanya seperti algoritma decision tree, naive bayes dan k-nearest neighbor. Adapun metode pengujian yang digunakan yaitu k-fold cross validation dan kurva ROC. Hasil metode data mining pada dataset kualitas air berupa komparasi performance dari ketiga algoritma tersebut sehingga akan didapatkan algoritma terbaik dalam mengklasifikasi kualitas air. Berdasarkan komparasi algoritma hasil pengujian semua metode yang digunakan maka dapatkan bahwa algoritma decesion tree menjadi algoritma yang paling akurat dalam mengklasifikasi data kualitas air dengan akurasi sebesar 94,94% dan nilai AUC sebesar 0,865 sehingga termasuk golongan klasifikasi yang baik.

***Keywords** : Algoritma, Data Mining, Komparasi, Kualitas Air*

1. PENDAHULUAN

Air merupakan sumber kehidupan terutama bagi kehidupan manusia. Sekitar 71% wilayah Bumi merupakan air [1]. Maka dari itu kualitas air tersebut wajib dijaga demi keberlangsungan kehidupan manusia dan alam sekitarnya. Kualitas air adalah suatu ukuran kondisi air dilihat dari karakteristik fisik, kimiawi, dan biologisnya [2]. Adanya pembangunan yang semakin pesat mendorong banyaknya penggunaan lahan di sepanjang aliran sungai. Hal ini bisa dilihat, terutama sungai-sungai di perkotaan, yang berubah fungsi menjadi permukiman dan kegiatan industri. Perubahan fungsi ini menimbulkan kekhawatiran akan menurunnya kualitas air yang mengalir sepanjang sungai tersebut.

Kondisi kualitas air yang tidak pernah dipermasalahkan oleh kebanyakan masyarakat sehingga tidak diketahui secara pasti kondisi ataupun kadar kualitas air tersebut untuk kesehatan [3]. Adapun beberapa faktor tercemarnya lingkungan akibat dari aktifitas industri seperti manufaktur, pertambangan, konstruksi, pembuangan limbah ke air sungai secara sembarangan dan masih banyak masyarakat yang belum sadar akan bahaya pembuangan sampah sembarangan ke sumber disekitar lingkungan air tersebut [3]. Penurunan kualitas air tidak hanya diakibatkan oleh limbah industri, tetapi juga diakibatkan oleh limbah rumah tangga baik limbah cair maupun limbah padat [4].

Pengecekan kualitas air merupakan salah satu upaya untuk mengkontrol jika ada penyakit dan bakteri pada air kolam terutama bagi pembudidaya ikan tertentu sehingga dapat dilakukan tindakan pencegahan apabila terjadi penurunan kualitas air [5]. Parameter kualitas air meliputi mikrobiologi, kimia anorganik, parameter fisik, serta parameter kimia. Parameter kualitas air diperlu diketahui karena air mengandung zat mineral yang terlarut didalam air, tetapi tidak semua zat-zat mineral yang terkandung didalam air dapat dikonsumsi dengan baik oleh tubuh manusia dikarenakan air juga sering terkena oleh bakteri maupun zat-zat mineral yang berbahaya bagi tubuh manusia.

Tujuan penelitian ini adalah menerapkan metode data mining untuk mengklasifikasi kualitas air berdasarkan parameter kualitas air meliputi mikrobiologi, kimia anorganik, dan parameter

kimia. Beberapa algoritma yang digunakan seperti algoritma *decision tree, naive bayes* dan *k-nearest neighbor*. Algoritma *decesion tree* adalah algoritma yang digunakan untuk menghasilkan sebuah pohon keputusan berdasarkan pemilihan atribut yang memiliki nilai *gain* tertinggi [6]. Adapun *naive bayes* adalah metode klasifikasi yang dapat memprediksi probabilitas sebuah class, sehingga dapat menghasilkan keputusan berdasarkan data pembelajaran dengan memberikan akurasi klasifikasi yang kompetitif dan efisiensi komputasi [7]. Sedangkan algortima k-nearest neighbor (KNN) adalah merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatnya [8]. Hasil metode data mining pada dataset kualitas air berupa komparasi performa dari ketiga algoritma tersebut sehingga akan didapatkan algoritma terbaik dalam mengklasifikasi kualitas air.

2. METODE PENELITIAN

Model penelitian yang digunakan dalam penelitian ini adalah menggunakan model penelitian eksperimen dan evaluasi. Penelitian ini bertujuan untuk mengkomparasi dan mengevaluasi algoritma klasifikasi data mining pada kasus kualitas air. Data yang digunakan dalam penelitian ini adalah data sekunder yang peneliti peroleh pada Kaggle datasets.

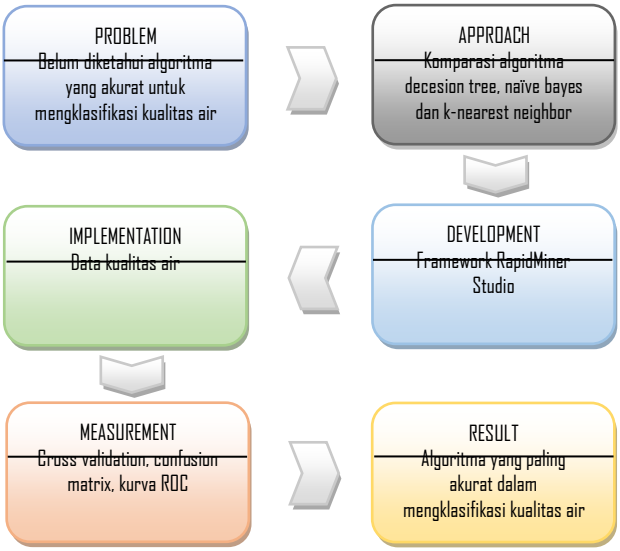
Tabel 1. Deskripsi Dataset Kualitas Air

Nama	Keterangan
Aluminium	berbahaya jika lebih besar dari 2,8
Amonia	berbahaya jika lebih besar dari 32,5
Arsenik	berbahaya jika lebih besar dari 0,01
Barium	berbahaya jika lebih besar dari 2
Kadmium	berbahaya jika lebih besar dari 0,005
Chloramine	berbahaya jika lebih besar dari 4
Kromium	berbahaya jika lebih besar dari 0,1
Tembaga	berbahaya jika lebih besar dari 1,3
Flouride	berbahaya jika lebih besar dari 1,5
Bakteri	berbahaya jika lebih besar dari 0
Virus	berbahaya jika lebih besar dari 0
Timbal	berbahaya jika lebih besar dari 0,015
Nitrat	berbahaya jika lebih besar dari 10
Nitrit	berbahaya jika lebih besar dari 1
Merkuri	berbahaya jika lebih besar dari 0,002
Perklorat	berbahaya jika lebih besar dari 56
Radium	berbahaya jika lebih besar dari 5
Selenium	berbahaya jika lebih besar dari 0,5
Perak	berbahaya jika lebih besar dari 0,1
Uranium	berbahaya jika lebih besar dari 0,3
Kualitas Air	atribut kelas {0 - tidak aman, 1 - aman}

Data yang telah dikumpulkan diidentifikasi, diseleksi, dibersihkan dan ditransformasikan ke bentuk yang diinginkan sehingga dapat dilakukan persiapan ketahap selanjutnya dalam pembuatan model.

Selanjutnya melakukan eksperimen dan pengujian model dengan pembagian data ke dalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.

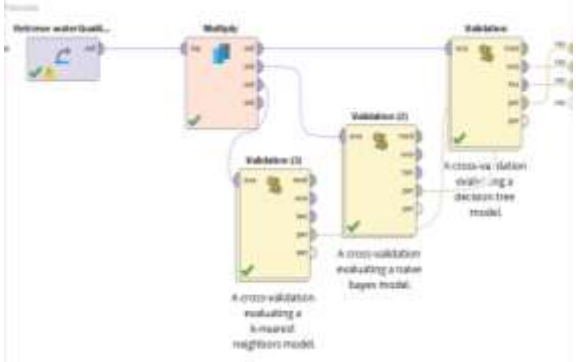
Kemudian untuk menguji model, pada penelitian ini, digunakan metode *cross validation*, *confusion matrix*, dan kurva ROC (*receiver operating characteristic*) [9].



Gambar 1. Kerangka Pemikiran

3. HASIL DAN PEMBAHASAN

Eksperimen dan pengujian model dengan pembagian data ke dalam data latihan (*training data*) dan data uji (*testing data*) menggunakan *k-fold cross validation* yang merupakan pilihan terbaik untuk mendapatkan hasil validasi yang akurat dengan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian.

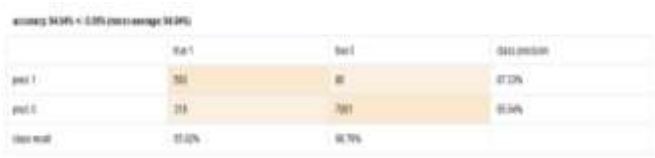


Gambar 2. Desain Cross Validation

Hasil pengujian *cross validation* didapatkan akurasi *decision tree* sebesar 94,94%, *naive bayes*

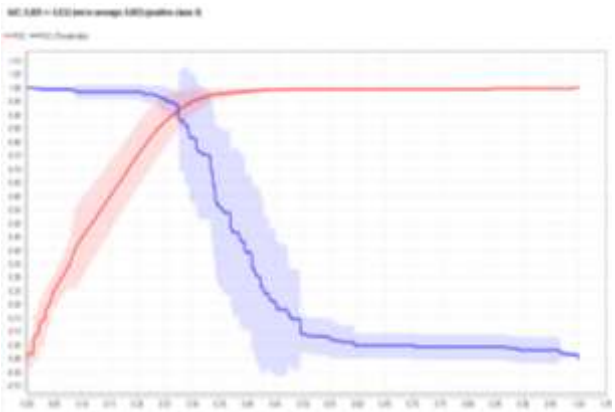
sebesar 84,79%, *k-nearest neighbor* sebesar 87,86%.

Berdasarkan *confusion matrix*, diketahui dari 7999 data, 593 diklasifikasikan aman sesuai dengan prediksi yang dilakukan dengan metode *decision tree*, lalu 86 data diprediksi aman tetapi ternyata termasuk tidak aman, 7001 data sesuai diprediksi tidak aman, dan 319 data diprediksi tidak aman ternyata aman.



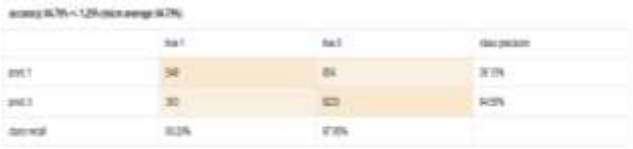
Gambar 3. Confusion Matrix Decision Tree

Gambar 4. Performance Vektor Decision Tree



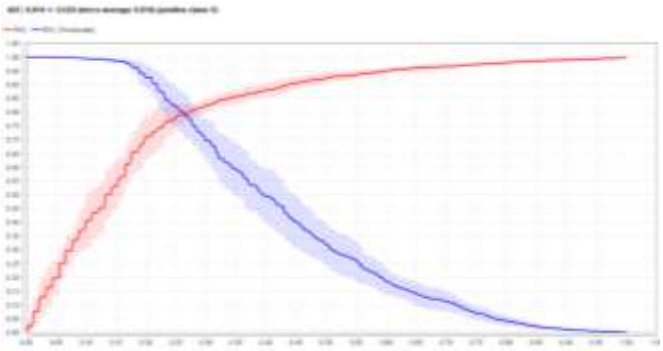
Gambar 5. Kurva ROC Decision Tree

Selanjutnya berdasarkan *confusion matrix*, diketahui dari 7999 data, 549 diklasifikasikan aman sesuai dengan prediksi yang dilakukan dengan metode *naive bayes*, lalu 854 data diprediksi aman tetapi ternyata termasuk tidak aman, 6233 data sesuai diprediksi tidak aman, dan 363 data diprediksi tidak aman ternyata aman.



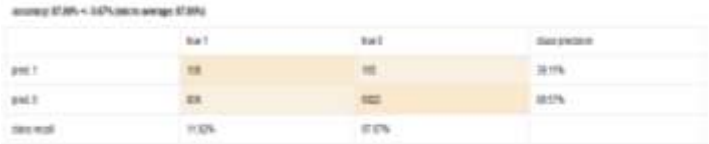
Gambar 6. Confusion Matrix Naive Bayes

Gambar 7. Performance Vektor Naive Bayes



Gambar 8. Kurva ROC Naive Bayes

Selanjutnya berdasarkan *confusion matrix*, diketahui dari 7999 data, 106 diklasifikasikan aman sesuai dengan prediksi yang dilakukan dengan metode *k-nearest neighbor*, lalu 165 data diprediksi aman tetapi ternyata termasuk tidak aman, 6922 data sesuai diprediksi tidak aman, dan 806 data diprediksi tidak aman ternyata aman.



Gambar 9. Confusion Matrix K-Nearest Neighbor

Gambar 10. Performance Vektor K-Nearest Neighbor

Gambar 11. Kurva ROC K-Nearest Neighbor

Berikut perbandingan *performance* akurasi dan nilai AUC dari ketiga algoritma yang diuji

Algoritma	Akurasi	Nilai AUC
Decision Tree	94,94%	0,865
Naive Bayes	84,79%.	0,814
K-NN	87,86%	0,725

Hasil komparasi uji akurasi algortima didapatkan bahwa algoritma *decision tree* menjadi algoritma yang paling akurat karena memiliki nilai akurasi yang paling tinggi jika dibandingkan dengan algoritma lainnya, diikuti algoritma *k-nearest neighbor* kemudian algoritma *naive bayes*.

Berdasarkan hasil uji kehandalan algoritma dari nilai AUC yang dihasilkan kurva ROC maka kehandalan algoritma dalam melakukan klasifikasi terbagi kedalam lima golongan yakni [10]:

Nilai Uji Kehandalan (AUC)	Klasifikasi
0,90 – 1,00	sangat baik
0,80 – 0,90	baik
0.70 – 0.80	cukup
0.60 – 0.70	Buruk
0.50 – 0.60	gagal

Hasil komparasi uji kehandalan algoritma maka didapatkan bahwa algoritma *decision tree* dan *naive bayes* termasuk kedalam golongan klasifikasi yang baik sedangkan algoritma *k-nearest neighbor* termasuk kedalam golongan klasifikasi cukup.

Berdasarkan hasil pengujian semua metode yang digunakan maka dapatkan bahwa algoritma *decision tree* menjadi algoritma yang paling akurat dalam mengklasifikasi data kualitas air.

4. KESIMPULAN

Dalam penelitian ini dilakukan pengujian model menggunakan algoritma *decision tree, naive bayes* dan *k-nearest neighbor* menggunakan dataset kualitas air. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling akurat dalam mengklasifikasi kualitas air. Pengukuran kinerja ketiga algoritma tersebut menggunakan metode pengujian *cross validation, confusion matrix* dan *kurva ROC*, diketahui bahwa algoritma *decision tree* memiliki nilai *accuracy* tertinggi sebesar 94,94% dan AUC sebesar 0,865 sehingga termasuk kedalam golongan klasifikasi yang baik. Dengan demikian, algoritma *decision tree* merupakan metode yang baik dalam pengklasifikasian data kualitas air dan dapat memberikan pemecahan untuk permasalahan penentuan kualitas air yang aman untuk dikonsumsi.

5. REFERENSI

[1] M. A. Rahman, N. Hidayat, and A. Afif Supianto, "Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Vol. 2, No. 12, Desember 2018, hlm. 6346-6353 e-ISSN*, vol. 2, no. 12, pp. 925–928, 2018.

[2] Wikipedia, "Kualitas Air," *Wikipedia.org*, 2022. https://id.wikipedia.org/wiki/Kualitas_air (accessed May 09, 2022).

[3] R. M. S. Tumangger, "Komparasi Metode Data Mining Support Vector Machine Dengan Naive Bayes Untuk Klasifikasi Status Kualitas Air," Universitas Brawijaya Malang, 2020.

[4] N. Mahsyar And E. R. Wijaya, "Analisis Kualitas Air Dan Metode Pengendalian Pencemaran Air Sungai Bangkala Kabupaten Jeneponto Oleh : Nurlina

Mahsyar Abstrak," Universitas Muhammadiyah Makassar, 2020.

[5] Y. T. K. Yunior and K. Kusrini, "Sistem Monitoring Kualitas Air Pada Budidaya Perikanan Berbasis IoT dan Manajemen Data," *Citec J.*, vol. 6, no. 2, p. 153, 2021, doi: 10.24076/citec.2019v6i2.251.

[6] M. R. Matondang, M. R. Lubis, and H. S. Tambunan, "Analisis Data mining dengan Metode C.45 pada Klasifikasi Kenaikan Rata-Rata Volume Perikanan Tangkap," *Brahmana J. Penerapan Kecerdasan Buatan*, vol. 2, no. 2, pp. 74–81, 2021, doi: 10.30645/brahmana.v2i2.68.

[7] H. Leidiyana, "Komparasi Algoritma Klasifikasi Data Mining Dalam Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor," Sekolah Tinggi Manajemen Informatika Dan Komputer Nusa Mandiri Jakarta, 2011.

[8] M. S. Mustafa and I. W. Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," *Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. VIII, no. 1, pp. 1–10, 2019.

[9] E. Karyadiputra and Z. Zaenuddin, "Penerapan Algoritma Decision Tree C4.5 Berbasis Seleksi Atribut Chi Squared Untuk Klasifikasi Tingkat Pengetahuan Ibu Dalam Pemberian Asi Eksklusif Pada Bayi," *Technol. J. Ilm.*, vol. 11, no. 1, p. 7, 2020, doi: 10.31602/tji.v11i1.2685.

[10] F. Gorunescu, *Data Mining : Concepts, Models and Techniques*, vol. 2, no. January 2013. 2015.