KLASIFIKASI KUALITAS AIR BERSIH DI JAKARTA MENGGUNAKAN ALGORITMA DECISION TREE DAN ALGORITMA NAÏVE BAYES

Endang Sri Palupi
Teknologi Informasi, Universitas Bina Sarana Informatika
Jalan Kramat Raya No. 98 Jakarta Pusat
endang.epl@bsi.ac.id

ABSTRAK

Kebutuhan air bersih sangatlah penting untuk kehidupan manusia. Air yang tercemar dapat berdampak buruk pada tubuh, seperti menimbulkan penyakit diare, kolera, disentri, tipes, cacingan, penyakit kulit hingga keracunan. Oleh sebab itu menggunakan dan menjaga kualitas air bersih sangatlah penting. Kebutuhan manusia akan air sangat kompleks antara lain untuk minum, masak, mandi, mencuci dan sebagainya. Menurut perhitungan WHO di Negara-negara maju setiap orang memerlukan air antara 60-120 liter per hari. Sementara di Negara-negara berkembang termasuk Indonesia setiap orang memerlukan air antara 30-60 liter per hari. Manfaat menjaga kualitas air adalah untuk melindungi kesehatan manusia, karena beberapa polutan menimbulkan risiko terhadap kesehatan manusia. Standar kualitas air melindungi kesehatan manusia dan menghindari biaya yang terkait dengan perawatan medis, hilangnya produktivitas, dan bahkan hilangnya nyawa. Penulis melakukan klasifikasi kualitas air bersih di Jakarta menggunakan algoritma *decision tree* dan algoritma *naïve bayes* dengan tujuan membuat aplikasi yang dapat digunakan untuk klasifikasi kualitas air bersih di Jakarta sehingga dapat mengelompokkan air yang layak dikonsumsi atau tidak. Hasil dari penelitian ini nilai akurasi menggunakan algoritma *naïve bayes* sebesar 72,48% dengan AUC sebesar 0.803 sedangkan nilai akurasi menggunakan algoritma *decision tree* 90,83% dengan AUC sebesar 0.861. Hasil penelitian menyimpulkan bahwa algoritma *naïve bayes* dan algortima *decision tree* dapat menjadi metode yang baik untuk klasifikasi dalam *data mining*.

Kata kunci : *decision tree, klasifikasi, naïve bayes*

1. PENDAHULUAN

Air yang berkualitas adalah air yang memenuhi baku mutu air minum yang ditetapkan oleh Peraturan Menteri Kesehatan RI No. 492/MENKES/PER/IV/2010 dimana air harus terbebas dari segala macam mikroorganisme yang patogen maupun apatogen dan bahan kimia berbahaya lainnya. [1] DKI Jakarta dilintasi oleh 13 sungai besar dan beberapa sungai kecil serta 40 situ tersebar di 5 wilayah kota yang sangat potensial sebagai air permukaan untuk menunjang kehidupan manusia. Dengan pertumbuhan penduduk DKI yang pesat dan perkembangan pemanfaatannya, ada kecenderungan terjadinya perubahan pada kondisi dan kualitas air sungai dan situ di DKI Jakarta.

Standar Air Bersih bertujuan untuk memastikan bahwa air yang dikonsumsi oleh masyarakat aman dan tidak membahayakan kesehatan manusia. Selain itu, Standar Air Bersih juga memiliki tujuan untuk melindungi lingkungan dan ekosistem dari pencemaran air yang disebabkan oleh kegiatan manusia.

Setiap negara memiliki Standar Air Bersih yang berbeda-beda, tergantung pada kebutuhan dan kondisi geografis dan lingkungan setempat. Pada umumnya, hal tersebut ditetapkan oleh pemerintah melalui badan pengatur lingkungan atau kesehatan. Standar Air Bersih terus diperbaharui dan ditingkatkan untuk menjaga kualitas air yang dikonsumsi oleh masyarakat. Perubahan-perubahan ini didasarkan pada penelitian dan pengembangan terbaru dalam bidang pengolahan air dan juga berkaitan dengan perkembangan teknologi dan industri yang dapat berdampak pada kualitas air.[2]

Berdasarkan data WHO, 19% penduduk dunia memiliki sumber air yang tidak aman. Selain itu 829.000 orang setiap tahun meninggal dikarenakan diare akibat air yang tidak aman dan sanitasi yang buruk. Berdasarkan data Bappenas tahun 2018 akses air minum layak di Indonesia adalah sebesar 87,75% dengan 6,8% adalah akses air minum aman. Penelitian menunjukkan bahwa ada hubungan antara penyediaan air minum dengan daya saing bangsa. Sumber air minum yang kurang menyebabkan daya saing yang rendah. Banyak masyarakat menghabiskan uangnya untuk berobat dan membeli air. Masyarakat yang sakit tentu produktivitasnya rendah. Indonesia juga menghadapi masalah kualitas air permukaan, di mana 52% sungai sudah tercemar berat. Jika hanya mengandalkan air permukaan tentu tantangannya besar, termasuk penyediaan teknologi pengolahan air. Oleh sebab itu, pemanfaatan air tanah sebagai sumber air baku tentu diperlukan dengan tetap memelihara air tanah itu sendiri karena air tanah adalah reservoir alami yang relatif gratis jika dibandingkan dengan reservoir buatan.[3]

Classification merupakan metode yang digunakan apabila atributnya berupa *numerik* atau *nominal*, namun labelnya harus berupa *nominal*. Pada metode *data mining* ini, dilakukan pengelompokan atau pengklasifikasian berdasarkan hubungan antara variabel kriteria dengan variabel target.[4]

Dalam melakukan metode klasifikasi, ada proses estimasi yang bernama simple/single split yaitu

memisahkan data untuk training (70%) dan testing (30%). Hal ini digunakan untuk melihat prediksi dari akurasi metode klasifikasi tersebut. Proses lainnya dalam metode klasifikasi adalah *k-Fold Cross Validation*, data dipisahkan dengan jumlah yang sama kedalam subsets kemudian dilakukan training/testing.[5]

Penelitian ini bertujuan untuk memberikan gambaran pada saat terkini mengenai kualitas air sungai dan situ yang ada di DKI Jakarta serta upaya pengendalian pencemaran air yang mungkin dapat dilakukan.[6] Mengingat pentingnya kualitas air bersih, penulis melakukan klasifikasi kualitas air bersih di Jakarta menggunakan algoritma *decision tree* dan algoritma *naïve bayes*. Maksud dari mengklasifikasi kualitas air bersih pada dasarnya untuk mengkontrol ada tidaknya penyakit, bakteri maupun kuman dalam air sehingga dapat mengelompokkan air yang layak dikonsumsi atau tidak.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Pada tahun 2023 Sutisna dan rekan melakukan penelitian klasifikasi kualitas air bersih menggunakan metode *naïve bayes*, hasil akurasi menggunakan *rapidminer* sebesar 97,35%, data menggunakan data sekunder diambil dari hasil wawancara dan observiasi dan data primer diambil dari studi pustaka dan *textbook*. Sedangkan penulis menggunakan metode 2 algoritma sebagai perbandingan yaitu algoritma *decision tree* dan algoritma *naïve bayes* pemodelan juga menggunakan *rapidminer* menggunakan *cross validation* dan *performance klasifikasi*, dan dataset diambil dari website : https://katalog.satudata.go.id/dataset/data-kualitas-air-sungai.[7]

Jurnal yang berjudul Penerapan Metode *Naïve Bayes* Untuk Mengetahui Kualitas Air Di Jakarta ditulis oleh Yunita Sartika Sari pada tahun 2021. Hasil penelitian nilai akurasi sebesar 50,6%, nilai akurasi dihitung secara manual dengan data yang didapatkan melalui wawancara, literatur pustaka, dan observasi. Sedangkan penelitian ini penulis menggunakan *framework rapidminer* dengan membandingkan 2 algoritma yaitu *naïve bayes* dan *decision tree*. Hasilnya nilai akurasi tertinggi yaitu algoritma *decision tree* sebesar 90.83% dengan data yang diambil dari website : https://katalog.satudata.go.id/dataset/data-kualitas-air-sungai. [8]

Bayu Prihambodo dan rekan pada tahun 2023 melakukan penelitian klasifikasi kualitas air sungai menggunakan algoritma *k-nearest neighbor* (K-NN), hasil penelitian memperoleh akurasi 78,46% dengan nilai data *training* 70% dan data *testing* 30% serta nilai K= 15, sedangkan penulis untuk *split* data menggunakan *cross validation* secara otomatis. Penelitian ini penulis juga menggunakan *framework rapidminer* studio sebagai analisis data dan mendapatkan output hasil klasifikasi. [9]

Penelitian yang berjudul Metode KNN untuk menentukan Kualitas Air dilakukan oleh Fahmi Malik, bertujuan untuk membuat aplikasi untuk menentukan kualitas air.Atribut yang digunakan dalam dataset adalah unsur yg terdapat dalam sampel air yang digunakan apakah layak dikonsumsi atau tidak, dataset diambil dari kaggle.com dengan judul "waterQuality1". Penelitian ini menggunakan MATLAB sedangkan penulis dalam penelitian ini menggunakan *framework rapidminer studio*.[10]

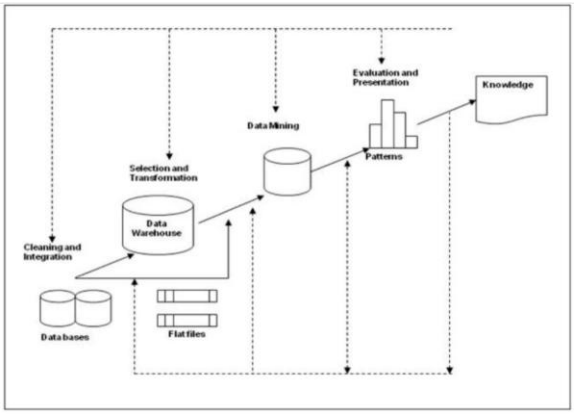
Implementasi algoritma *naïve bayes* untuk memprediksi kualitas air yang dapat dikonsumsi telah dilakukan oleh Adrian Wisnu Saputra dan rekan pada tahun 2024.Hasil akurasi sebesar 65,08% dengan presisi 62,08 % dan recall 26,47%. Penelitian ini juga menggunakan *rapidminer* dan untuk dataset diambil dari kaggle.com dengan 24 attribute dan 1 label. Sedangkan penulis mengambil data dari website https://katalog.satudata.go.id/dataset/data-kualitas-air-sungai dengan 150 record, 27 variabel dan 1 label. Penulis dalam penelitian ini menggunakan *cross validation* sebagai salah satu teknik yang sangat penting dalam evaluasi dan pengujian model *machine learning*. [11]

Jurnal yang ditulis oleh GL Pritalia pada tahun 2022 melakukan analisis komparatif menggunakan beberapa algoritma yang ada dalam *machine learning* untuk mengklasifikasi air layak minum. Hasil penelitian tersebut algoritma paling optimal adalah *random forest* mencapai 87% dan akurasi test 85% dengan menggunakan 7 fitur terbaik, sedangkan tingkat *miss rate* paling kecil 15% yaitu algoritma *decision tree*. Sedangkan penulis dalam penelitian ini melakukan komparasi akurasi dan AUC hanya dengan 2 algoritma yaitu *decision tree* dan *naïve bayes* dengan hasil akurasi tertinggi *decision tree* sebesar 90,83% sedangkan algoritma *naïve bayes* sebesar 72,48%.[12]

2.2. *Data Mining*

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk *mengekstraksi* dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. *Data mining* adalah sebuah proses pencarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Teknik *data mining* digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Istilah lain yang sering dikaitkan dengan *data mining* diantaranya *knowledge discovery (mining) in databases, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, dan business intelligence. Data mining* adalah bagian integral dari *Knowledge Discovery in Databases* (KDD). Keseluruhan proses KDD untuk *konversi raw* data

(data mentah) ke dalam informasi yang berguna ditunjukkan dalam Gambar 1. [13]



Gambar 1. Tahapan *KDD Data Mining*

2.3. *Decision Tree*

Decision tree merupakan salah satu cara *data processing* dalam memprediksi masa depan dengan cara membangun klasifikasi atau *regresi model* dalam bentuk struktur pohon. Hal tersebut dilakukan dengan cara memecah terus ke dalam himpunan bagian yang lebih kecil lalu pada saat itu juga sebuah pohon keputusan secara bertahap dikembangkan. Hasil akhir dari proses tersebut adalah pohon dengan node keputusan dan node daun. Sebuah node keputusan (misalnya, Cuaca/ Outlook) memiliki dua atau lebih cabang (misalnya, Panas, Berawan dan Hujan). Decision Tree juga berguna untuk dieksplorasi data, menemukan hubungan antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan *eksplorasi* data dan pemodelan yang salah langkah pertama yang sangat baik dalam proses pemodelan yang digunakan sebagai model akhir untuk beberapa teknik lainnya. Kelebihan lain dari metode ini adalah mampu mengeliminasi perhitungan atau data-data yang tidak diperlukan. Karena sampel yang ada biasanya hanya diuji berdasarkan kriteria atau kelas tertentu. [14]

2.4. Naïve Bayes

Naïve bayesian classifier mengasumsikan bahwa keberadaan sebuah atribut *(variabel*) tidak ada kaitannya dengan beradaan atribut yang lain karena asumsi atribut tidak saling terkait. Adapun cara kerja dari proses perhitungan *naïve bayes* yaitu tahapan diawali dengan mengambil data testing, menghitung nilai *probabilitas* setiap kriteria berdasarkan dari data latih, setelah menghitung nilai probabilitas setiap kriteria berdasarkan dari data latih, selanjutnya menghitung nilai *probabilitas* tiap-tiap *fitur* berdasarkan data testing dan data latih.

Mengalikan hasil dari P(Y) pada masing- masing kelas dan data uji.[15] *Naïve Bayes Classifier* yaitu salah satu metode machine learning yang menggunakan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memperkirakan probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. [16]

2.5. RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*) yang diciptakan dengan menggunakan bahasa pemrograman java sehingga bisa diakses oleh semua sistem operasi. *RapidMiner* dapat dijadikan solusi dalam melakukan analisis terhadap datamining, dengan menggunakan teknik deskriptif dan prediksi yang diberikan kepada pengguna sehingga dapat mengambil keputusan yang paling baik. *RapidMiner* yang digunakan pada penelitian ini adalah *RapidMiner* versi Studio 10.1.3. [17]

RapidMiner adalah sebuah aplikasi atau *Sofware* peragkat lunak yang berfungsi sebagai alat pembelajan pada ilmu data mining. *platfrom* dikembangkan oleh suatu perusahaan yang bertujuan untuk bisnis komersial, penelitian, pendidikan, pelatihan, serta semua langkah dalam pembelajaran yang menyangkut pada suatu data yang besar. *RapidMiner* telah membuktikan telah sukses mencapai untuk memberikan dampak bisnis yang cepat bagi lebih dari 40.000 organisasi di setiap industri untuk mendorong pendapatan, dan menghindari resiko. Tata Kerja *RapidMiner* sudah terbukti memberikan 3 tahapan yang mempuni, diantaranya :

a. Transparansi penuh & tata kelola untuk pembelajaran mesin yang satu diantaranya memiliki peran *Easy to Tune* yang berarti buka, periksa dan konfigurasi Ulang dari 1.500 + blok bangunan visual algoritma *machine learning* dan operator ilmu data.
b. Mudah dijelaskan *platform* otomatis membangun alur kerja analitik visual yang dikisahkan untuk berkolaborasi dengan pemangku kepentingan bisnis dan menawarkan metode yang kaya untuk penjelasan model.
c. *End-to-end* merupakan hal yang mengenai tentang mengubah data dari sumber apapun, mengoptimasikan pemilihan dan validasi model *machine learning* terbaik. [18]

3. METODE PENELITIAN

3.1. Jenis Penelitian

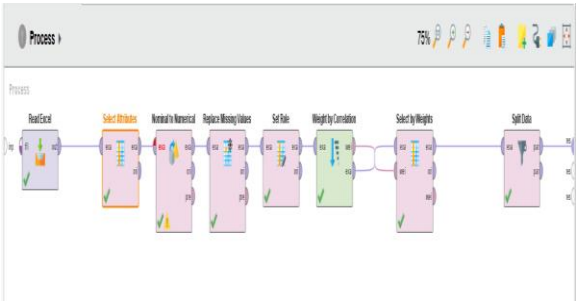
Penelitian ini menggunakan metode *kuantitatif deskriptif* yaitu konsisten dengan variabel penelitian, fokus pada permasalahan aktual dan fenomena yang sedang terjadi, serta menyajikan hasil penelitian dalam bentuk angka-angka yang bermakna. [19] mendeskripsikan, meneliti, dan menjelaskan sesuatu yang dipelajari apa adanya, dan menarik kesimpulan dari fenomena yang dapat diamati dengan menggunakan angka-angka. Penelitian *deskriptif kuantitatif* adalah penelitian yang menggambarkan variabel secara apa adanya didukung dengan data-data berupa angka yang dihasilkan dari keadaan sebenarnya. Penelitian deskriptif dapat bersifat kuantitatif karena mengumpulkan data yang dapat

diukur untuk menganalisis sampel populasi secara statistik . Angka-angka ini dapat menunjukkan pola, hubungan, dan tren dari waktu ke waktu dan dapat ditemukan menggunakan survei, jajak pendapat, dan eksperimen. Metode penelitian *deskriptif kuantitatif* adalah suatu metode yang bertujuan untuk membuat gambar atau *deskriptif* tentang suatu keadaan secara objektif yang menggunakan angka, mulai dari pengumpulan data, penafsiran terhadap data tersebut serta penampilan dan hasilnya. [20]

Tahapan penelitian terlihat pada gambar 3 berikut penjelasannya :

a. Pengambilan data
   Data diambil dari website : https://katalog.satudata.go.id/dataset/data-kualitas-air-sungai. Dataset sebanyak 2015 record, dengan 9 atribute dan 1 label.

b. *Preprocessing data*
   Data *preprocessing* adalah tahap untuk melakukan sebuah proses awal dalam pengolahan data. Pada tahap ini data yang akan diolah bertujuan untuk menghindarkan dari data yang mengganggu (noise) atau data yang tidak konsisten.[21] Tahapan untuk menghilangkan beberapa permasalahan yang bisa mengganggu saat pemrosesan data. Hal tersebut karena banyak data yang formatnya tidak konsisten. Data *preprocessing* merupakan teknik paling awal sebelum melakukan *data mining*. Berikut tampilan *preprocessing* dalam *RapidMiner* pada gambar 2.



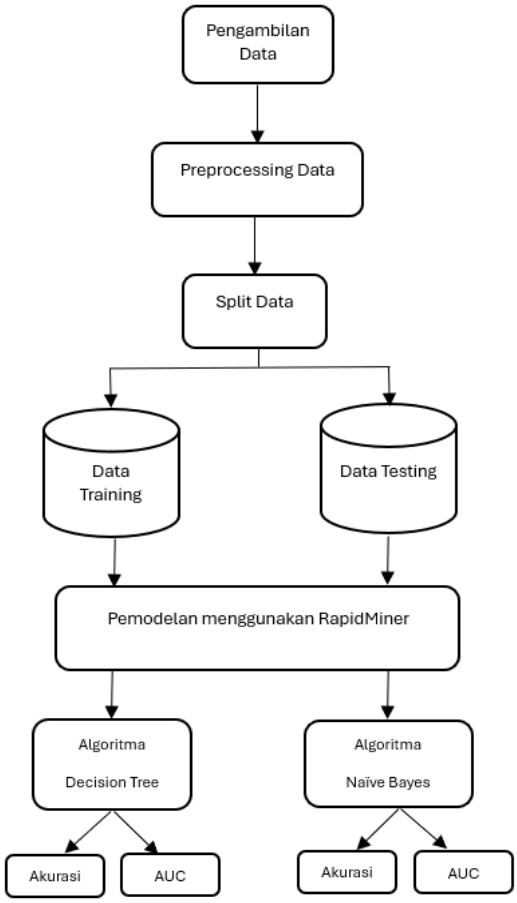
Gambar 2. *Preprocessing Data*

c. *Split* data; data training dan data testing
   Tujuannya untuk memastikan bahwa model tidak hanya berfokus pada pembelajaran dari kumpulan data tertentu, tetapi juga berkinerja baik saat berhadapan dengan data baru dan tak terlihat. Pada penelitian ini penulis menggunakan *feature cross validation*.

d. Pemodelan menggunakan *RapidMiner*
   Pemodelan menggunakan *RapidMiner* versi Studio 10.1.3, menggunakan *cross validation* dan *confusion matrix*.

e. Algoritma *naïve bayes*
   Menghasilkan nilai akurasi dan AUC

f. Algoritman *desicion tree*
   Menghasilkan nilai akurasi dan AUC



Gambar 3. Tahapan Penelitian

4. HASIL DAN PEMBAHASAN

4.1. Preprocessing Data

Berikut *dataset* yang digunakan dalam penelitian ini pada file berikut data-kualitas-air-sungai.csv. Sebelum data diolah menggunakan algoritma, penulis melakukan *preprocessing data* untuk mengurangi *attribute* yang nilainya kurang dari 0,15 dengan menggunakan *feature select-by-weight. Weight by Correlation* merupakan salah satu metode yang digunakan untuk meningkatkan nilai akurasi dari model machine learning yang dikembangkan. Metode pemilihan fitur yang dipilih adalah *Weight by Correlation*, yaitu melakukan pembobotan atribut dengan cara menghubungkan (mengkorelasikan) satu atribut dengan atribut lainnya.[22]

Berikut tabel 1 adalah hasil dari pembobotan, yang diwarnai kuning adalah *attribute* dengan nilai kurang dari 0,15.

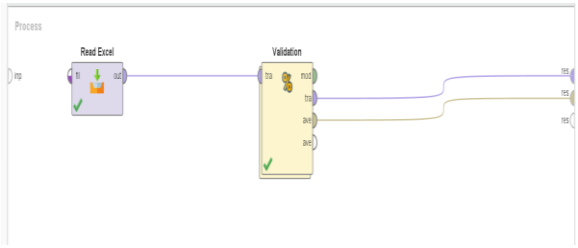
Tabel 1. Hasil Pembobotan *attribute*

Attribute	Weight
pH	0,515
Suhu_(0C)	0,080
DHL_(µS/cm)	0,211
Salinitas_(ppt)	0,486
Do_(mg/L)	0,096
TDS_(mg/L)	0,325
TSS_(mg/L)	0,355
BOD_(mg/L)	0,411

Attribute	Weight
COD_(mg/L)	0.078
Nitrat_(mg/L)	0,419
Nitrit_(mg/L)	0,330
Fenol_(µg/L)	0,026
Sulfida_(mg/L)	0,324
Sianida_(mg/L)	0,324
Phospat_(mg/L)	0,410
MBAS_(µg/L)	0,215
M&L_(µg/L)	0,413
NH3N_(mg/L)	0,255
Klorin_(mg/L)	0,321
EColi_(MPN/100ml)	0,415
Coliform_(MPN/100ml)	0,212
Mercury_(mg/L)	0,319
Nilai_IP	0,402
Status_MutuAir	0,303

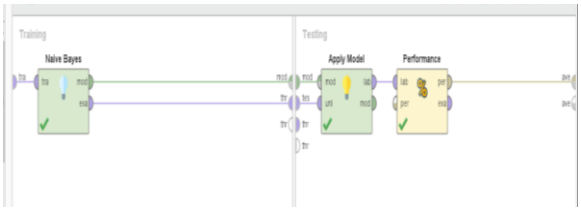
4.2. Pemodelan Algoritma *Naïve Bayes*

Pemodelan menggunakan algoritma *naïve bayes* dengan *cross validation* menggunakan *RapidMiner* terlihat pada gambar 4.

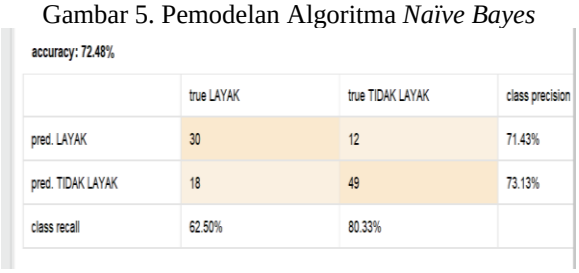


Gambar 4. Pemodelan *RapidMiner*

Gambar 5 adalah pemodelan menggunakan algoritma *naïve bayes* dengan *apply model* dan *performance* untuk klasifikasi.



Gambar 5. Pemodelan Algoritma *Naïve Bayes*

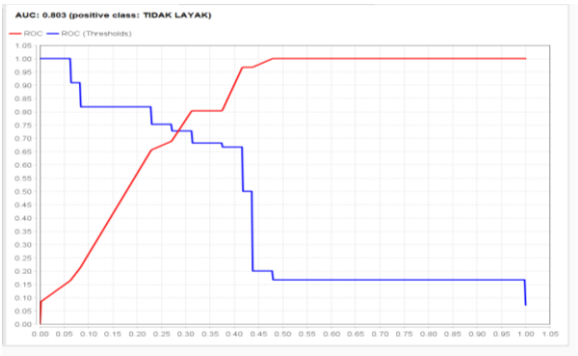


accuracy: 72.48%

	true LAYAK	true TIDAK LAYAK	class precision
pred. LAYAK	30	12	71.43%
pred. TIDAK LAYAK	18	49	73.13%
class recall	62.50%	80.33%	

Gambar 6. Hasil Akurasi Algoritma *Naïve Bayes*

Gambar 6 adalah hasil akurasi pemodelan menggunakan algoritma *naïve bayes* sebesar 72,48% dengan *class recall* 62,505 LAYAK dan *class precision* 73,13 TIDAK LAYAK termasuk ke dalam klasifikasi yang baik.



Gambar 7. Hasil AUC Algoritma *Naïve Bayes*

Gambar 7 adalah hasil AUC (*Area Under Curve*) sebesar 0.803, dengan hasil ini hasil klasifikasi kualitas air ini termasuk dalam kategori klasifikasi yang baik.

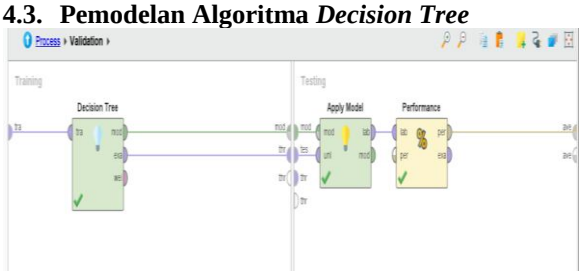
Perhitungan akurasi algoritma *naïve bayes* :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$Accuracy = \frac{30+49}{30+49+18+12} \times 100\%$$

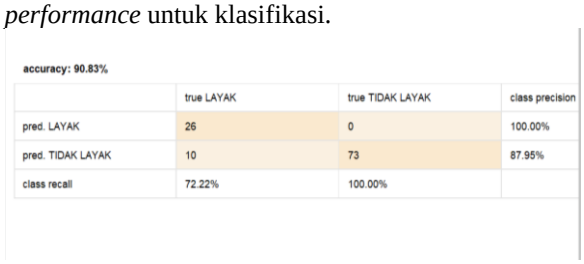
$$Accuracy = \frac{79}{109} \times 100\% = 0{,}7248 = 72.48\%$$

4.3. Pemodelan Algoritma *Decision Tree*



Gambar 8. Pemodelan Algoritma *Decision Tree*

Gambar 8 adalah pemodelan menggunakan algoritma *decision tree* dengan *apply model* dan *performance* untuk klasifikasi.



accuracy: 90.83%

	true LAYAK	true TIDAK LAYAK	class precision
pred. LAYAK	26	0	100.00%
pred. TIDAK LAYAK	10	73	87.95%
class recall	72.22%	100.00%	

Gambar 9. Hasil Akurasi Algoritma *Decision Tree*

Gambar 9 adalah hasil akurasi pemodelan menggunakan algoritma *decision tree* sebesar 90,83% dengan *class recall* 72,22% LAYAK dan *class precision* 87,95% TIDAK LAYAK termasuk ke dalam klasifikasi yang baik.

Gambar 10. Hasil AUC Algoritma *Decision Tree*

Gambar 10 adalah hasil AUC (*Area Under Curve*) sebesar 0.861, dengan hasil ini hasil klasifikasi kualitas air ini termasuk dalam kategori klasifikasi yang baik.

Perhitungan akurasi algoritma *decision tree* :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$Accuracy = \frac{26+73}{26+73+10+0} \times 100\%$$

$$Accuracy = \frac{99}{109} \times 100\% = 0{,}9083 = 90.83\%$$

Berikut tabel 2 adalah hasil komparasi algoritma naïve bayes dan algoritma decision tree :

Tabel. 2 Komparasi Akurasi dan AUC

Algoritma	Akurasi	AUC
Naïve Bayes	72.48%	0.803
Decision Tree	90.83%	0.861

5. KESIMPULAN DAN SARAN

Penelitian ini menghasilkan klasifikasi kualitas air yang layak atau tidak layak dikonsumsi. Dalam pemodelan algoritma, *split* data *training* dan data *testing* dilakukan secara otomatis menggunakan *feature cross validation* pada *RapidMiner*. Sebelum proses pemodelan algoritma, dilakukan *preprocessing* data dengan menghilangkan data yang tidak perlu dan mengurangi atribut yang nilainya paling rendah sehingga data siap diolah.Algoritma *decision tree* memiliki nilai akurasi yang lebih baik dengan nilai akurasi 90.83% dan AUC 0.861 sedangkan algoritma *naïve bayes* memiliki nilai akurasi 72.48% dan AUC 0.803. Algoritma *decision tree* dan *naïve bayes* bisa dijadikan sebagai klasifikasi yang cukup baik dalam *data mining*.Untuk penelitian selanjutnya bisa menggunakan data yang lebih update, dengan sampling yang lebih banyak, dan menggunakan *optimisasi* pada algoritma *data mining* sehingga bisa mendapatkan hasil yang lebih baik.

DAFTAR PUSTAKA

[1] C. David Laksamana and E. Prasetyo, "Analisis Kualitas Fisik Air Desa Cranggang Kecamatan Dawe Kabupaten Kudus," *J. Kesehat. Masy.*, vol. 5 (1), pp. 26–35, 2017.

[2] A. Herdiasa, "Standar Air Bersih: Pentingnya Menjaga Kualitas Air yang Kita Konsumsi," *PDAM Info*, 2023. https://pdaminfo.pdampintar.id/blog/lainnya/standar-air-bersih-pentingnya-menjaga-kualitas-air-yang-kita-konsumsi

[3] Adi Permana, "Urgensi Menjaga Ketersediaan Air Bersih yang Aman di Indonesia," *itb.ac.id*, 2020. https://itb.ac.id/berita/urgensi-menjaga-ketersediaan-air-bersih-yang-aman-di-indonesia/57576

[4] Acer Indonesia, "Apa Itu Metode Data Mining? Ini Klasifikasi dan Contohnya!," 2023. https://www.acerid.com/berita-bisnis/metode-data-mining-dan-contohnya

[5] Angelia Fransiska, "METODE DATA MINING CLASSIFICATION," *binus ac.id*, 2021. https://sis.binus.ac.id/2021/10/22/metode-data-mining-classification/

[6] D. Hendrawan, "Kualitas Air Sungai Dan Situ Di DKI Jakarta," *MAKARA Technol. Ser.*, vol. 9, no. 1, pp. 13–19, 2010, doi: 10.7454/mst.v9i1.315.

[7] Sutisna and N. M. Yuniar, "Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 243–246, 2023, [Online]. Available: https://doi.org/10.55338/saintek.v5i1.1383

[8] Y. S. Sari, "Penerapan Metode Naïve Bayes Untuk Mengetahui Kualitas Air Di Jakarta," *J. Ilm. FIFO*, vol. 13, no. 2, p. 222, 2021, doi: 10.22441/fifo.2021.v13i2.010.

[9] B. Prihambodo, A. W. F. Y, E. Prayoga, and A. Jaffar, "Klasifikasi Kualitas Air Sungai Berbasis Teknik Data Mining Dengan Metode K-Nearest Neighbor (K-NN)," *Emit. J. Tek. Elektro*, vol. 23 No 1, pp. 31–36, 2023, [Online]. Available: https://journals2.ums.ac.id/index.php/emitor/index.

[10] F. M. N. Akbar, "Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air," *J. Tekno Kompak*, vol. 18, no. 1, p. 28, 2024, doi: 10.33365/jtk.v18i1.3241.

[11] A. Wisnu Saputra, A. Irma Purnamasari, and I. Ali, "Implementasi Algoritma Naïve Bayes Untuk Memprediksi Kualitas Air Yang Dapat Di Konsumsi," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 133–140, 2024, doi: 10.36040/jati.v8i1.8292.

[12] Generosa Lukhayu Pritalia, "Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 43–55, 2022, doi: 10.24002/konstelasi.v2i1.5630.

[13] Anggada Maulana, "Konsep Dasar Data Mining," in *Konsep Data Mining*, vol. 1, 2018, pp. 1–16.

[14] Ardi Satyo Ramadhan, "Decision Tree Algoritma Beserta Contohnya Pada Data Mining," *binus ac.id*, 2022. https://sis.binus.ac.id/2022/01/21/decision-tree-algoritma-beserta-contohnya-pada-data-mining/

[15] E. Martantoh and N. Yanih, "Implementasi Metode Naïve Bayes Untuk Klasifikasi Karakteristik Kepribadiaan Siswa Di Sekolah MTS Darussa'adah Menggunakan Php Mysql," *J. Teknol. Sist. Inf.*, vol. 3, no. 2, pp. 166–175, 2022, doi: 10.35957/jtsi.v3i2.2896.

[16] A. Z. Macfud, A. P. Kusuma, and W. D. Puspitasari, "Analisis Algoritma Naïve Bayes Classifier (NBC) Pada Klasifikasi Tingkat Minat Barang Di Toko Violet Cell," vol. 7, no. 1, pp. 87–94, 2023.

[17] M. Ridwan, L. Bahrudi, and E. Damanik, "Penerapan Algoritma C5.0 Dalam Menentukan Tingkat Pemahaman Mahasiswa Terhadap Pembelajaran Daring," *KOMPUTA J. Ilm. Komput. dan Inform.*, vol. 11, no. 1, 2022, doi: 10.34010/komputa.v11i1.7386.

[18] Admin, "RapidMiner Tren Aplikasi Pengolahan Data Science," *The Education University*, 2022. https://tve.upi.edu/rapidminer-tren-aplikasi-pengolahan-data-science/

[19] Sugiyono, *Metode Penelitian Kuantitatif*. Bandung: Alfabeta, 2018.

[20] A. Kunto, *Prosedure Penelitian Suatu Pendekatan Praktek*. Jakarta: PT Rineka Cipta, 2006.

[21] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 9, no. 02, pp. 75–81, 2021, doi: 10.33884/jif.v9i02.3755.

[22] I. Santoso, W. Gata, and A. B. Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 5–11, 2021.