KLASIFIKASI KUALITAS AIR MINUM MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHTBOR*

SKRIPSI

Oleh :
QANITA FARAH FADILAH
NIM. 19650139

PROGRAM STUDI TEKNIK INFROMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

KLASIFIKASI KUALITAS AIR MINUM MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHTBOR*

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
QANITA FARAH FADILAH
NIM. 19650139

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

HALAMAN PERSETUJUAN

KLASIFIKASI KUALITAS AIR MINUM MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR*

SKRIPSI

Oleh :
QANITA FARAH FADILAH
NIM. 19650139

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 18 Juni 2025

Pembimbing I,	Pembimbing II,
Dr. Ir. M. Amin Hariyadi, M.T	Dr. Irwan Budi Santoso, M.Kom
NIP. 19670018 200501 1 001	NIP. 19770103 201101 1 004

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

KLASIFIKASI KUALITAS AIR MINUM MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

SKRIPSI

Oleh :
QANITA FARAH FADILAH
NIM. 19650139

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 2 Juli 2025

Susunan Dewan Penguji

Ketua Penguji	: Okta Qomaruddin Aziz, M.Kom NIP. 19911019 201903 1 013	( )
Anggota Penguji I	: Khadijah F.H. Holle, M.Kom NIP. 19900626 202203 2 002	( )
Anggota Penguji II	: Dr. Ir. M. Amin Hariyadi, M.T NIP. 19670018 200501 1 001	( )
Anggota Penguji III	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Qanita Farah Fadilah
NIM	: 19650139
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Klasifikasi Kualitas Air Minum Menggunakan Algoritma K-Nearest Neighbor

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 2 Juli 2025
Yang membuat pernyataan,

Qanita Farah Fadilah
NIM. 19650139

MOTTO

Yakin sama takdir Allah Swt. dan tetap berbaik sangka kepada-Nya

HALAMAN PERSEMBAHAN

Alhamdulillaahirobbil'aalamiin segala puji bagi Allah subhanahu wa ta'ala, serta shalawat dan salam kepada Rasulullah Muhammad SAW. Penulis mempersembahkan hasil karya ini kepada :

Papa, mama, dan keluarga yang telah menjadi motivasi terbesar bagi penulis yang selalu mendukung, mendoakan, dan memberikan limpahan kasih sayang yang tak terhingga.

Dan seluruh teman-teman tercinta yang telah memberikan banyak *support*, perhatian, kasih sayang serta arahan sehingga penulis dapat menyelesaikan skripsi ini.

Dan juga diri sendiri yang telah berusaha dan berjuang sampai sejauh ini.

KATA PENGANTAR

Segala puji dan syukur kehadirat Allah SWT, karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi ini dengan judul "Klasifikasi Kualitas Air Minum Menggunakan Algoritma *K-Nearest Neighbor*". Skripsi ini disusun untuk memenuhi salah satu syarat guna memperoleh gelar Sarjana Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Dalam proses penyusunan skripsi ini, penulis banyak mendapatkan bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini, penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Harini, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Ir. Fachrul Kurniawan, M.MT., IPU., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. M. Amin Hariyadi, M.T selaku Dosen Pembimbing 1 dan Dr. Irwan Budi Santoso, M.Kom selaku Dosen Pembimbing 2 yang membimbing penulis dengan penuh kesabaran dan keikhlasan dalam meluangkan waktunya untuk membimbing sehingga penulis dapat menyelesaikan tugas akhir ini.
5. Okta Qomaruddin Aziz, M.Kom selaku dosen penguji 1 dan Khadijah F. H. Holle, M.Kom selaku dosen penguji 2 yang telah memberikan arahan dalam menyelesaikan skripsi ini.

6. Ajib Hanani, M.T selaku Wali Dosen yang senantiasa memberikan dukungan, arahan, bimbingan, serta motivasi selama pendidikan.
7. Seluruh Dosen dan Staf di Program Studi Teknik Informatika, dengan ikhlas memberikan ilmu, bantuan, serta dorongan semangat selama perkuliahan.
8. Kedua orang tua penulis, Sugeng Hariyadi dan Sri Hartati, kedua kakak perempuan saya, kakak laki-laki saya, dan kakak ipar saya yang telah memberikan banyak dukungan, doa serta bantuan sehingga penulis mampu menyelesaikan masa studi hingga mencapai gelar sarjana.
9. Serta satu orang yang berarti bagi saya Rahim Ma'ruf yang senantiasa meluangkan waktu untuk memberikan dorongan, motivasi, hingga materi, dalam masa perkuliahan hingga proses penyelesaian skripsi.
10. Teman-teman terdekat saya, yang selalu memberikan semangat, bantuan, dan kebersamaan selama masa studi.
11. Seluruh pihak yang telah terlibat secara langsung maupun tidak langsung dalam membantu menyelesaikan skripsi ini.
12. Diri saya sendiri, yang telah mampu melawan rasa malas dan bekerja keras untuk menyelesaikan skripsi ini. Terimakasih telah mampu kooperatif dan bertahan dalam menikmati proses pengerjaan skripsi.

Akhir kata, semoga skripsi ini dapat memberikan manfaat bagi semua pihak yang berkepentingan dan dapat menambah wawasan serta pengetahuan di bidang ilmu Teknik Informatika.

Malang, 22 Juni 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Fadilah, Qanita Farah. 2025. **Klasifikasi Kualitas Air Minum Menggunakan K-Nearest Neighbor.** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. M. Amin Hariyadi, M.T (II) Dr. Irwan Budi Santoso, M.Kom.

.

Kata kunci: Kualitas Air, Klasifikasi, K-Nearest Neighbor.

Kualitas air yang aman untuk dikonsumsi merupakan faktor penting bagi kesehatan masyarakat. Sistem yang dapat mengklasifikasikan kualitas air minum secara cepat dan akurat sangat diperlukan guna mencegah tercemarnya air yang dapat menimbulkan berbagai penyakit. Penelitian ini dilakukan untuk mengklasifikasikan kualitas air yang layak dikonsumsi atau tidak menggunakan algoritma *K-Nearest Neighbor* berdasarkan parameter fisika dan kimia seperti pH, kekeruhan, TDS, dan kandungan sulfat. Algoritma ini digunakan dalam menghitung probabilitas kualitas air yang layak atau tidak layak dikonsumsi. Data yang digunakan dalam penelitian ini diambil dari website *Kaggle* sebanyak 3.276 data yang akan diproses melalui tahap *preprocessing* data, normalisasi, penyeimbangan data, serta penerapan algoritma KNN dengan perbandingan rasio data dan jumlah nilai k. Hasil evaluasi dari model yang dihasilkan diukur menggunakan tabel *Confusion Matrix* yang menghasilkan tingkat akurasi tertinggi 68%, presisi 73%, *recall* 63%, dan *f1-score* 67% dengan nilai k (tetangga terdekat) = 3 dan rasio perbandingan data 90:10. Hasil akurasi tersebut menunjukkan bahwa metode KNN cukup efektif dalam mengklasifikasikan kualitas air minum.

ABSTRACT

Fadilah, Qanita Farah. 2025. **Drinking Water Quality Classification Using K-Nearest Neighbor.** Undergraduate Thesis. Department of Informatics Engineering Faculty of Science and Technology Maulana Malik Ibrahim State Islamic University Malang. Supervisor: (I) Dr. Ir. M. Amin Hariyadi, M.T (II) Dr. Irwan Budi Santoso, M.Kom.

.

Keywords: Water Quality, Classification, K-Nearest Neighbor.

The quality of water that is safe for consumption is an important factor for public health. A system that can classify the quality of drinking water quickly and accurately is needed to prevent water contamination that can cause various diseases. This research was conducted to classify the quality of water that is suitable for consumption or not using the K-Nearest Neighbor algorithm based on physical and chemical parameters such as pH, turbidity, TDS, and sulfate content. This algorithm is used in calculating the probability of water quality that is suitable or not suitable for consumption. The data used in this study were taken from the Kaggle website as much as 3,276 data which will be processed through the stages of data preprocessing, normalization, data balancing, and application of the KNN algorithm with a comparison of data ratios and the number of k values. The evaluation results of the resulting model are measured using the Confusion Matrix table which produces the highest accuracy rate of 68%, precision 73%, recall 63%, and f1-score 67% with a value of k (nearest neighbor) = 3 and a data comparison ratio of 90:10. The accuracy results show that the KNN method is quite effective in classifying drinking water quality.

صلختسم ثحبلا

فاضله، قانيتا فرح. ٢٠٢٥. **تصنيف جودة مياه الشرب باستخدام نظام ك–أقرب جار**. رسالة جامعية. قسم الهندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية، مالانج. المشرفون: (١) الدوكتور إروان بودي سانتوسو، ماجستير في العلوم والتكنولوجيا (٢) الدوكتورالمهندس محمد أمين هريدي ، ماجستير في العلوم والتكنولوجيا

كلمة أساسية : جودة المياه، التصنيف، ك–أقرب جار

جودة المياه الصالحة للاستهلاك عاملٌ مهمٌّ للصحة العامة، ويتطلب الأمر نظامًا قادرًا على تصنيف جودة مياه الشرب بسرعة ودقة لمنع تلوث المياه الذي قد يُسبب أمراضًا مُختلفة. أُجري هذا البحث لتصنيف جودة المياه الصالحة للاستهلاك أو غير الصالحة باستخدام خوارزمية "ك–أقرب جار" بناءً على المعايير الفيزيائية والكيميائية مثل الرقم الهيدروجيني، والعكارة، والمواد الصلبة الذائبة، ومحتوى الكبريتات. يتم استخدام هذه الخوارزمية لحساب احتمالية كون جودة المياه مناسبة أو غير مناسبة للاستهلاك. تم أخذ البيانات المستخدمة في هذه الدراسة من موقع Kaggle الإلكتروني بما يصل إلى ٣.٢٧٦ بيانات والتي سيتم معالجتها من خلال مراحل معالجة البيانات المسبقة، والتطبيع، وموازنة البيانات، وتطبيق خوارزمية ك–أقرب جار مع مقارنة نسب البيانات وعدد قيم ك. تم قياس نتائج تقييم النموذج الناتج باستخدام جدول مصفوفة الارتباك الذي أنتج أعلى مستوى دقة بنسبة ٦٨، ودقة بنسبة ٧٣٪، واسترجاع بنسبة ٦٣٪، ونتيجة ف١ بنسبة ٦٧٪ بقيمة ك (أقرب جار) = ٣ ونسبة مقارنة البيانات ٩٠:١٠. وتظهر نتائج الدقة أن طريقة ك–أقرب جار فعالة جدًا في تصنيف جودة مياه الشرب.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dari tahun ke tahun kebutuhan air bersih semakin bertambah. Daerah resapan yang sempit terutama di wilayah perkotaan dan semakin banyaknya yang tidak mempertimbangkan keseimbangan lingkungan menyebabkan ketersediaan air bersih menjadi sangat terbatas. Sehingga ketersediaan sumber air bersih menjadi sangat berkurang. Menurunnya ketersediaan air dan kualitas air bersih di daerah Indonesia disebabkan banyaknya air yang tercemar berasal dari sampah sisa limbah domestik dan rumah tangga. Kualitas air yang buruk mengakibatkan timbulnya berbagai penyakit seperti penyakit diare.

Saat ini, pencemaran air membutuhkan perhatian khusus karena sudah menjadi masalah global. Menurut WHO, 785 juta orang masih belum memiliki akses terhadap air minum yang aman dan layak konsumsi. Bahkan, diperkirakan 2 milliar orang meminum air yang sudah terkontaminasi dengan kotoran (Nada Osseiran & Yemi Lufadeju, 2019). Di Indonesia, isu pencemaran air sudah lama terjadi. Pencemaran di Indonesia penyebab utamanya berasal dari limbah cucian piring, kotoran manusia, hewan, dan pupuk (Na'imah, 2021).

Data mining yang juga dikenal sebagai KDD (*Knowledge Discovery in Databases*), merupakan proses analisis terhadap himpunan data dalam skala besar untuk mengidentifikasi pola atau hubungan yang bermakna (Harahap & Sulindawaty, 2020). Secara public, data mining adalah metode pencarian pola-pola

pengetahuan yang tidak diketahui sebelumnya dan tersembunyi berdasarkan dari suatu data *warehouse*, sekumpulan data yang besar pada database, atau media penyimpanan lainnya. Hasil dari proses *data mining* dapat dimanfaatkan sebagai dasar dalam melakukan evaluasi dan perbaikan terhadap pengambilan keputusan di masa mendatang.

Data mining mencakup berbagai teknik dalam proses identifikasi pola tersembunyi, salah satunya adalah teknik klasifikasi. Teknik ini secara luas diaplikasikan untuk memprediksi nilai berdasarkan sejumlah atribut, serta berperan sebagai metode pembelajaran data yang bertujuan menentukan kelas dari label tertentu. Teknik dalam klasifikasi yaitu membangun model atau mengklasifikasikan data berdasarkan training set dan nilai-nilai pada label kelas dalam mengklasifikasikan atribut tertentu. Terdapat beberapa algoritma dalam teknik klasifikasi, namun pada penelitian ini menggunakan algoritma k-Nearest Neighbor. Menurut Setianto et al., (2019) K-Nearest Neighbor (KNN) adalah salah satu metode klasifikasi yang bekerja dengan membandingkan data *training* dan menentukan kelas suatu data berdasarkan kedekatan jaraknya terhadap sejumlah tetangga terdekat yang ditentukan oleh nilai *k*. Algoritma K-Nearest Neighbor (KNN) juga memiliki beberapa karakteristik diantaranya yaitu sebagai algortima yang bersifat *lazy learning* dan *non-parametric*. Berdasarkan kedua sifat tersebut menjadikan KNN sebagai algoritma yang mampu menangani berbagai jenis data dan algoritma yang sangat fleksibel.

Pada Al-Quran terdapat ayat Al-Qur'an dalam ayat 172 pada Surah Al-Baqarah yang menyampaikan firman dari Allah SWT:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا كُلُوْا مِنْ طَيِّبٰتِ مَا رَزَقْنٰكُمْ وَاشْكُرُوْا لِلّٰهِ اِنْ كُنْتُمْ اِيَّاهُ تَعْبُدُوْنَ

"*Wahai orang-orang yang beriman, makanlah dari rezeki yang baik yang kami berikan kepada kamu dan bersyukurlah kepada Allah Swt, jika kamu hanya menyembah-Nya.*" *(QS. Al-Baqarah : 172).*

Dalam Tafsir Al-Azhar, Buya Hamka memberi penjelasan bahwa Allah Swt. mengajak orang-orang beriman untuk hanya mengonsumsi makanan yang berkualitas baik (thayyib) dan halal, karena apapun yang dikonsumsi manusia akan berpengaruh kepada sikap dan budi pekertinya. Karena itu, Allah SWT menurunkan ayat tersebut sebagai petunjuk tentang jenis makanan yang layak dan baik untuk dikonsumsi seperti tumbuh-tumbuhan, buah-buahan, dan binatang ternak (Samsudin, 2020).

Dari dalil tersebut, kita sebagai umat Islam dianjurkan untuk mengonsumsi makanan dan minuman yang tidak hanya halal, tapi juga baik dan menyehatkan (thayyib). Thayyib yang dimaksud adalah makanan atau minuman yang bersih. Oleh karena itu, kita dianjurkan untuk berusaha bagaimana mendapatkan air yang berkualitas dengan tujuan air tersebut baik untuk dikonsumsi. Karena apa saja yang dikonsumsi manusia akan berpengaruh terhadap jiwa dan budi pekertinya. Pentingnya untuk mencari makanan atau minuman yang bersih dan berkualitas agar sesuatu yang masuk di dalam tubuh manusia akan menjadi baik menjalankan ibadah hanya untuk Allah SWT.

Salah satu penelitian sebelumnya dilakukan oleh (Rachmat et al., 2020) yang menguji kelayakan air minum menggunakan metode *K-Nearest Neighbor* di PDAM Tirta Raharja. Dari penelitian tersebut, diperoleh akurasi sebesar 93% untuk

data dengan label “Layak diminum” dan 98% untuk label “Tidak layak diminum”. Metode KNN yang digunakan memakai nilai k = 14, yang diperoleh melalui teknik *K-Fold Cross Validation* denagn total 1.818 data. Data penelitian ini bersumber dari hasil klasifikasi laboratorium PDAM Tirta Raharja, yang dikumpulkan lewat observasi dan wawancara. Penelitian lain dilakukan oleh (Tangkelayuk & Evangs, 2022) yang membandingkan kinerja algoritma KNN, *Decision Tree,* dan *Naïve Bayes* dalam proses klasifikasi data. Mereka menggunakan dataset *Water Quality* dari situs Kaggle dan ingin mengetahui metode mana yang paling akurat. Hasilnya, KNN memberikan performa terbaik dengan akurasi 86,88%, diikuti 80,84% oada *Decision Tree*, dan *Naïve Bayes* dengan akurasi 63,60%.

Berdasarkan penelitian tersebut, maka akan dilakukan klasifikasi pada kelayakan kualitas air minum dengan metode KNN. Algoritma KNN dipilih karena memanfaatkan kelebihan yang ada pada algoritma ini. Menurut (Jaelani, 2020) algoritma KNN memiliki beberapa kelebihan yaitu algoritma KNN sangat *non-linier*, efektif apabila data *training sample*-nya besar, mudah dipahami dan diterapkan, memiliki konsistensi yang kuat, serta proses penentuan parameter model lebih sederhana dan sedikit. Namun algoritma KNN cenderung sensitif jika data terdapat *missing value* dan *outlier* sehingga perlu pemrosesan data yang cermat. Selain itu, KNN juga kurang optimal untuk digunakan pada data berdimensi tinggi. Kendati demikian, berdasarkan karakteristik serta kelebihan dan keterbatasan metode ini, KNN dinilai sesuai untuk digunakan pada dataset dalam penelitian ini. Dataset yang dimaksud bersumber dari situs Kaggle dan diupload oleh Aditya Kadiwal pada tahun 2021 yang berjudul *Water Quality* (*Drinking water*

potability). Data yang akan digunakan berjumlah 3.276 data. Hasil pada data tersebut diberikan label pada atribut *potability* yang menunjukkan apakah air tersebut aman untuk dikonsumsi manusia.

1.2 Identifikasi Masalah

Pada penelitian ini, identiikasi masalah berdasarkan latar belakang yaitu berapa besar tingkat akurasi yang dihasilkan dari algoritma KNN pada klasifikasi kualitas air?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk melihat seberapa akurat sistem dalam melakukan klasifikasi data algoritma KNN dalam melakukan pengukuran kualitas air pada data *Water Quality*.

1.4 Batasan Masalah

Batasan masalah pada klasifikasi kualitas air minum ini yaitu:

1. Kumpulan dataset pada klasifikasi kualitas air ini bersifat universal dengan total data berjumlah 3.276 *record* data.
2. Sumber data didapat dari *web* resmi "*Kaggle*".
3. Dataset yang digunakan terdiri dari 10 atribut, yaitu pH, *Hardness* (Kekerasan), *Solids* (Padatan), *Chloramines* (Kloramin), *Sulfate* (Sulfat), *Conductivity* (Konduktivitas), *Organic Carbon* (Karbon Organik), Trihalomethanes, *Turbidity* (Kekeruhan), dan *Potability* (sifat dapat diminum).

1.5 Manfaat Penelitian

Berdasarkan hasil penelitian ini, diharapkan hasil yang diperoleh dapat memberikan kontribusi positif bagi masyarakat dalam mengevaluasi kualitas air minum serta membantu dalam menentukan kelayakan air untuk dikonsumsi.

BAB II

STUDI PUSTAKA

2.1 Penelitian Tekait

Penelitian sebelumnya oleh (Said et al., 2022) mengkaji kelayakan air minum dengan menerapkan algoritma *K-Nearest Neighbors* (KNN). Dataset yang digunakan diambil dari situs Kaggle, dan evaluasi model dilakukan menggunakan *confusion matrix* untuk memperoleh tingkat akurasi. Hasil pengujian menunjukkan bahwa model menghasilkan akurasi sebesar 85,24% dengan penggunaan nilai *k* sebanyak 3 sebagai parameter jumlah tetangga terdekat.

Penelitian terdahulu lainnya dilakukan oleh (Tangkelayuk & Evangs, 2022) yang dalam penelitiannya menerapkan KNN, *Naïve Bayes,* dan *Decision Tree* untuk melakukan perbandingan dengan ketiga metode tersebut. Penelitian ini mencari metode yang akurasinya tertinggi, dimana *Dataset Water Quality* diambil dari Kaggle. Dalam penelitian ini, algoritma *K-Nearest Neighbors* menunjukkan performa paling baik dibandingkan metode lainnya. KNN berhasil mencapai akurasi tertinggi sebesar 86,88%, sedangkan algoritma *Naïve Bayes* hanya memperoleh akurasi 63,60%, dan *Decision Tree* mendapatkan hasil sebesar 80,84%.

Penelitian selanjutnya dilakukan oleh (Vidiastanta et al., 2020) dengan melakukan komparasi antara algoritma SVM (*Support Vector Machine*) dan KNN (*K-Nearest Neighbor*). Hasil dari penelitian ini yaitu akurasi tertinggi pada algoritma KNN sebesar 88.94%, sedangkan algoritma SVM akurasinya 87,71%.

Namun disebabkan karena sifat data yang masih *imbalanced*, hasil akurasi masih terdapat kesalahan klasifikasi pada data dengan nilai SVM sebesar 12,29% dan KNN hanya sebesar 11,06%.

Penelitian terkait lainnya dilakukan sebelumnya dilakukan oleh (Nurmahaludin & Cahyono, 2019) menggunakan metode KNN dan *K-Means*. Peneliti tersebut melakukan komparasi dengan kedua algoritma tersebut yang menghasilkan nilai akurasi terbaik pada algoritma KNN sebesar 88% dibandingkan dengan *K-Means* yaitu 76% berdasarkan data yang diambil dari PDAM daerah Banjarmasin.

Penelitian terdahulu yang pernah dilakukan oleh (Malik & Akbar, 2024) menggunakan metode *K-Nearest Neighbor* dalam menentukan kualitas air dengan mempertimbangkan 20 unsur kimia yang relevan untuk membuat aplikasi yang lebih inovatif. Penelitian tersebut melibatkan penggunaan dataset sekunder untuk mengkategorikan data pada kualitas air dalam menerapkan algoritma KNN. Data kualitas air tersebut terdiri dari 20 unsur yang terdapat dalam 8000 dataset yang diimplementasikan menggunakan GUI MATLAB. Hasil pada penelitian tersebut adalah sebuah aplikasi GUI untuk menentukan tingkat keamanan kualitas air, dimana dengan metode KNN mampu menjadi alat yang efektif dalam mengklasifikasikan data kualitas air.

Penelitian terdahulu lainnya juga dilakukan oleh (Prihambodo et al., 2023) untuk mengklasifikasikan air sungai berdasarkan kualitasnya dengan metode *K-Nearst Neighbor*. Penelitian ini menggunakan data kualitas air sungai sebanyak 216 sampel yang dikelompokkan ke dalam empat kelas. Setiap sampel dianalisis

berdasarkan tujuh parameter yang termasuk dalam Indeks Kualitas Air (IKA). Hasil dari penelitian ini yaitu metode KNN dengan rasio data 70:30 dan akurasi sebesar 78,46% serta nilai k = 15. Data tersebut diperoleh dari Perusahaan Daerah Air Minum (PDAM) Kota Malang dan berisi 216 baris data yang merepresentasikan kondisi air di wilayah tersebut. Pada prosesnya menggunakan *software RapidMiner* yang sebelumnya melakukan validasi data sebagai *pre-prosessing* untuk memilah data.

Penelitian terdahulu didapatkan dari pengujian pada air minum yang layak ataupun tidak menggunakan *K-Nearest Neighbor* (KNN) oleh (Rachmat et al., 2020) di PDAM Tirta Raharja. Penelitian ini menghasilkan akurasi sebesar 93% untuk prediksi air yang termasuk kategori layak minum, dan 98% untuk kategori tidak layak minum. Pengujian dilakukan menggunakan metode *K-Nearest Neighbor* (KNN) dengan nilai k = 14, yang diperoleh melalui proses *K-Fold Cross Validation* dari total 1.818 data. Data yang digunakan dalam pengujian ini berasal dari hasil klasifikasi laboratorium milik PDAM Tirta Raharja, yang diperoleh melalui proses wawancara langsung dan observasi di lapangan.

Penelitian yang dilakukan oleh (Muhammad et al., 2020) bertujuan membangun sebuah sistem pemantauan kualitas air pada bendungan dengan menerapkan algoritma klasifikasi *k-Nearest Neighbors* (k-NN). Sistem tersebut dilengkapi dengan berbagai sensor, meliputi sensor kedalaman air, sensor suhu, sensor kekeruhan (*turbidity*), sensor pH, serta sensor *Total Dissolved Solid* (TDS), yang berfungsi untuk mengumpulkan data secara real-time sebagai dasar klasifikasi kualitas air. Dalam pengujian terdapat 3 lokasi bendungan yang dipilih yaitu desa Giringan, desa Golang, dan desa Kresek yang terletak di Kabupaten Madiun. Hasil yang

diperoleh yaitu *k*=5 dengan akurasi sebesar 92%, dimana pengujian dilakukan dengan 50 data latih dan 25 data uji. Data klasifikasi diambil dari 3 lokasi bendungan tersebut di Kabupaten Madiun yang terdiri dari 75 data. Hasil sistem yang didapatkan sebesar 4135 *ms* dan dilakukan pengujian waktu komputasi sebanyak 10 kali. Hasil klasifikasi terdapat 3 kelas yaitu "Bagus", "Sedang", dan "Jelek".

Penelitian terkait oleh (Sopiatul Ulum et al., 2023) membandingkan performa algoritma *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM) dalam proses klasifikasi kelayakan air minum. Dataset yang digunakan adalah *Water Potability* yang bersumber dari situs Kaggle, terdiri atas 3.276 entri data dan memuat sepuluh atribut, yaitu pH, *Sulfate, Conductivity, Hardness, Turbidity, Solids, Chloramines, Organic_carbon*, Trihalomethanes, dan *Potability*. Berdasarkan hasil pengujian, algoritma SVM menghasilkan akurasi tertinggi sebesar 69,764%, sedangkan KNN memperoleh akurasi sebesar 65,341%.

Penelitian terdahulu lainnya pernah dilakukan oleh (Zamri et al., 2022) yang melakukan perbandingan akurasi pada metode *Support Vector Machine*, *K-Nearest Neighbors*, dan *Decision Tree* untuk menentukan klasifikasi kualitas sungai yang paling akurat. Hasil yang didapatkan pada penelitian tersebut adalah metode KNN yang menggunakan perhitungan jarak *Euclidean distance* memiliki akurasi sebesar 81.48%, KNN dengan perhitungan *Hamming distance* 90.12%, dan KNN dengan menggunakan Entropy Estimator yang memiliki akurasi tertinggi dengan akurasi sebesar 99.90% dibandingkan dengan metode *Decision Tree* sebesar 97.53%, dan *Support Vector Machin*e sebesar 83.95%. Kumpulan data yang digunakan diambil

dari Sungai Trengganu, Malaysia. Dataset tersebut terdiri dari 405 sampel dengan 27 fitur yang terdapat 5 tingkatan level pencemaran sungai yaitu sangat bersih, bersih, tercemar ringan, tercemar, dan tercemar tinggi.

Penelitian yang dilakukan oleh (Mardiana et al., 2022) menggunakan metode Analisis Diskriminan yang divalidasi dengan teknik *K-Fold Cross Validation*, menggunakan 42 data sampel primer. Penelitian ini memfokuskan pada dua kategori sebagai variabel dependen, yaitu air yang tercemar ringan dan tercemar sedang, yang diperoleh dari perhitungan indeks pencemaran. Dari beberapa percobaan yang dilakukan, model terbaik ditemukan pada uji kedua, dengan rumus $D_2 = -3{,}500 + 0{,}087x_2$ dan tingkat kesalahan prediksi (APER) paling rendah sebesar 0,21.

Selanjutnya, penelitian terkait yang pernah dilakukan oleh (Mutoffar & Fadillah, 2022) yaitu melakukan klasifikasi dalam menentukan kualitas air tanah menggunakan *Random Forest* yang dijadikan sebuah Solusi untuk menangani air yang tercemar. Disebabkan di Provinsi DKI Jakarta memiliki masalah banjir yang sering terjadi membuat kualitas air sumur atau air tanah di wilayah tertentu menjadi tidak layak dikonsumsi. Data yang digunakan sebanyak 267 data dengan data *testing* sebanyak 53 data dan data *training* sebanyak 214 data. Data ini merupakan hasil analisa laboratorium air sumur di wilayah Jakarta. Hasil pada peniltin inin yaitu nilai sensitivitas sebesar 0.83 dan presisi sebesar 0.823, hasil dari klasifikasi *Random Forest* yaitu sebesar 82% dari 83% data.

Penelitian terkait lainnya dilakukan oleh (Sutisna & Yuniar, 2023) yang melakukan penelitian dengan metode *Naïve Bayes* untuk memprediksi kualitas air

yang jernih atau tidak jernih dengan data berjumlah 226 data dari suatu perusahaan dengan atribut ph, total khlor, dan *turbidity*. Tujuan dari penelitian ini yaitu mengetahui ciri dari kualitas air yang baik serta mengklasifikasikan kualitas air yang bersih, yang metodenya dengan melakukan wawancara dan observasi. Penelitian ini dilakukan menggunakan *tools RapidMiner*. Hasil dari klasifikasi ini adalah akurasi sebesar 97,35%.

Penelitian selanjutnya mengenai klasifikasi kualitas air menggunakan Metode ELM (*Extreme Learning Machine*) yang dilakukan oleh (Kualitas & Menggunakan, n.d.). Penelitian ini dilakukan untuk mengklasifikasi air yang layak diminum atau tidak diperoleh dari sumber air tawar. Data diperoleh dari Web Kaggle kelasnya masih belum seimbang (*imbalance data*) sehingga diperlukan metode SMOTE untuk menyeimbangkan kelas data. Haasil pada uji coba penelitian ini yaitu *node hidden* 100 dan *k-fold* 10 dengan tingkat akurasi 97.873%, spesifisitas 98.6408% dan sensitifitas 96.4789%.

Penelitian yang dilakukan oleh (Hartanti & Pradana, 2023) melakukan komparasi antara algoritma *Decission Tree*, *Logistic Regression*, SVM, dan ANN dalam identifikasi kualitas air dengan *Dataset* berjumlah 3276 data dengan tujuan untuk mendapatkan akurasi tertinggi. Data yang digunakan yaitu terdiri dari 9 Fitur atau Atribut (*Solids, Chloramines, PH, Hardness, Organic Carbon, Sulfate, Conductivity, Trihalomethanes, Turbidity*) dan 1 kelas sebagai label yaitu *Potability*. Hasil pada penelitian ini yaitu metode SVM sebesar 68,59%, *Decission Tree* sebesar 60,19%, metode ANN sebesar 69,54%, dan *Logistic Regression*

sebesar 62,80%. Berdasarkan beberapa penelitian terkait tersebut, maka dapat ditunjukkan pada Tabel 2.1 Penelitian Terdahulu.

Tabel 2.1 Penelitian Terdahulu

No.	Sumber	Data	Metode	Hasil
1.	(Said et al., 2022)	Data diambil dari situs *Kaggle*	algoritma *K-Nearest Neighbor* (Tingkat akurasi 85,24%)	Metode yang digunakan pada penelitian ini adalah algoritma *K-Nearest Neighbor*
2.	(Tangkelayuk & Evangs, 2022)	*Dataset Water Quality* yang diperoleh dari situs *kaggle*	metode KNN (sebesar 86.88%), *Naïve Bayes* (sebesar 63.60%.) dan *Decision Tree* (sebesar 80.84%)	
3.	(Vidiastanta et al., 2020)	Data kualitas air juga dikumpulkan dari hasil analisis laboratorium milik PDAM Kota Malang.	Algoritma *Support Vector Machine* (SVM sebesar 87,71%) dan *K-Nearest Neighbor* (sebesar 88.94%)	
4.	(Nurmahaludin & Cahyono, 2019)	Dari PDAM Bandarmasih yang berlokasi di Kota Banjarmasin.	algoritma *K-Nearest Neighbor* (akurasi sebesar 88%) dan *K-Means* (akurasi sebesar 76%.)	
5.	(Malik & Akbar, 2024)	Dataset *qualityWatter1* diperoleh dari situs Kaggle.com	Metode *K-Nearest Neighbor*	
6.	(Prihambodo et al., 2023)	PDAM Kota Malang dengan 216 *record* data	algoritme *K-Nearst Neighbor* (akurasi sebesar 78,46%)	
7.	(Rachmat et al., 2020)	data hasil klasifikasi Laboratorium PDAM Tirta Raharja	*K-Nearest Neighbor* (KNN) (akurasi 93%)	
8.	(Muhammad et al., 2020)	Data dari 3 lokasi bendungan di Kabupaten Madiun yang terdiri dari 75 data.	Metode *K-Nearest Neighbor* (KNN) (akurasi sebesar 92%)	
9.	(Sopiatul Ulum et al., 2023)	Data dari Kaggle yaitu dataset *water potability* dengan jumlah 3276 data	*K-Nearest Neighbors* (akurasi sebesar 65,341%) dan *Support Vector Machine* (akurasi sebesar 69,764%)	
10.	(Zamri et al., 2022)	Data dari Sungai Trengganu, Malaysia (405 sampel)	KNN (akurasi sebesar 99.90%), *Decision Tree* (sebesar 97.53%), dan *Support Vector Machine* (sebesar 83.95%)	
11.	(Mardiana et al., 2022)	Data primer 42 sampel air di kota Pontianak	Metode Analisis Diskriminan dengan *K Fold Cross Validation*	

No.	Sumber	Data	Metode	Hasil
12.	(Mutoffar & Fadillah, 2022)	Data dari hasil Analisa laboratorium air sumur di Provinsi DKI Jakarta sebanyak 267 data	*Random Forest* (nilai presisi 0.823 dan nilai sensitivitas 0.83)	
13.	(Sutisna & Yuniar, 2023)	Data dari suatu Perusahaan berjumlah 226 *record* dengan 8 atribut	Metode *Naïve Bayes* dengan tools RapidMiner (akurasi sebesar 97,35%)	
14.	(Kualitas & Menggunakan, n.d.)	Data dari kaggle pada 18 April 2023 sebanyak 20 parameter dengan 8000 total data	Metode *Extreme Learning Machine* (ELM) dengan menerapkan SMOTE untuk *imbalance* data (k-fold = 10 dengan nilai akurasi 97.8723%)	
15.	(Hartanti & Pradana, 2023)	Data publik kualitas air berjumlah 3276 baris dengan 9 Atribut	Metode *Decission Tree* (60,19%), metode *Logistic Regression* (62,80%), metode SVM (68,59%), dan metode ANN (69,54%)	

2.2 Kualitas Air

Berdasarkan Peraturan Menteri Kesehatan Nomor 492 Tahun 2010, air minum adalah air yang aman dikonsumsi langsung dan memenuhi standar kesehatan, baik setelah melalui proses pengolahan maupun tanpa pengolahan. Agar dianggap layak untuk diminum, air tersebut harus memenuhi kriteria tertentu yang mencakup aspek biologi, fisik, kimia, dan juga bebas dari zat radioaktif. Hal ini dikarenakan masyarakat dihimbau agar mengonsumsi air yang tidak menyebabkan dampak yang berbahaya bagi kesehatan. Prioritas utama dalam mencegah adanya penyakit pada air yang akan dikonsumsi yaitu memperbaiki kualitas air. Karena itulah, penting untuk melakukan langkah-langkah pencegahan agar kita tidak terkena penyakit yang bisa muncul akibat penggunaan air yang sudah tercemar, baik oleh bahan kimia maupun bakteri. Menjaga kualitas air sangat berpengaruh

terhadap kesehatan, karena air yang tidak bersih bisa menjadi sumber berbagai masalah kesehatan. (Addzikri & Rosariawari, 2023).

Air baku adalah air yang digunakan sebagai bahan dasar untuk diolah sesuai dengan kebutuhan tertentu, terutama untuk air minum. Kegunaan utamanya memang lebih banyak difokuskan pada kebutuhan konsumsi. Menurut Peraturan Pemerintah Nomor 16 Tahun 2005 tentang Sistem Penyediaan Air Minum, sumber air baku untuk air minum bisa berasal dari air tanah, air permukaan, atau air hujan, asalkan sudah memenuhi standar kualitas tertentu. Di antara semua sumber tersebut, sungai, danau, dan waduk termasuk yang paling bisa diandalkan karena kapasitasnya besar dan pasokannya cukup stabil. Saat ini, sebagian besar air baku untuk air minum di Indonesia memang masih berasal dari sumber-sumber air permukaan seperti sungai, danau, dan waduk.

Berdasarkan Undang-Undang Nomor 7 Tahun 2004 tentang Sumber Daya Air, pemerintah pusat dan daerah memiliki tanggung jawab dalam mengembangkan sistem penyediaan air minum. Tujuan utamanya adalah untuk meningkatkan kesejahteraan masyarakat, dengan memastikan bahwa kebutuhan dasar akan air minum dapat terpenuhi, baik dari segi jumlah, mutu, maupun ketersediaannya secara terus-menerus. (Addzikri & Rosariawari, 2023). Aturan mengenai standar kualitas air minum sudah dijelaskan dalam Peraturan Menteri Kesehatan Nomor 492 Tahun 2010. Selain itu, Undang-Undang Nomor 32 Tahun 2004 tentang Pemerintahan Daerah juga menegaskan bahwa penyediaan air minum merupakan tanggung jawab Pemerintah Kabupaten atau Kota. Dalam hal ini, Perusahaan Daerah Air Minum (PDAM) bertugas sebagai pelaksana layanan, sedangkan

pemerintah daerah berperan sebagai pengatur kebijakan agar masyarakat bisa mendapatkan akses air minum yang layak.

Sekarang ini, banyak sumber air yang sudah tercemar karena ulah manusia sendiri, seperti membuang sampah sembarangan atau limbah pabrik yang langsung dibuang ke sungai atau laut. Pencemaran air dapat berdampak pada kehidupan manusia yang berpengaruh pada kesehatannya. Jika terpapar dalam jangka waktu yang panjang, maka dampaknya akan sangat berbahaya. Oleh karena itu, kualitas air jadi sangat penting, mengingat air bukan termasuk sumber daya yang bisa diperbarui dengan mudah. Berikut beberapa resiko akibat perncemaran air terhadap kesehatan manusia :

a. Diare
b. Penyakit kulit
c. Demam berdarah
d. Disentri
e. *Trachoma* (infeksi mata)
f. Hepatitis
g. Kanker pada kulit, kandung kemih, dan paru-paru (Lobo, 2022).

2.2.1 Baku Mutu Air

Setiap jenis sumber air berpotensi dimanfaatkan sebagai air baku untuk keperluan air minum, dengan catatan bahwa semakin tinggi kualitas air baku tersebut, maka semakin layak pula untuk diolah menjadi air minum (Musdalifah, 2022). Menurut Peraturan Pemerintah Republik Indonesia Nomor 82 Tahun 2001 mengenai Pengendalian Pencemaran Air dan Pengelolaan Kualitas Air, pembagian

kelas mutu air dibagi menjadi 4 kelas berdasarkan tingkatan mutu air yang baik dan kemungkinan kegunaannya. Klasifikasi mutu air yang ditetapkan Pemerintah yaitu:

a. Kelas I: Air yang bisa langsung dipakai sebagai bahan baku air minum atau harus memenuhi standar mutu yang setara dengan air minum.
b. Kelas II : Air yang dimanfaatkan untuk kegiatan seperti tempat rekreasi air, budidaya ikan air tawar, peternakan, pengairan tanaman, atau kebutuhan lain yang serupa.
c. Kelas III : Air yang digunakan untuk budidaya ikan air tawar, irigasi tanaman, peternakan, dan keperluan lain yang membutuhkan kualitas air sesuai dengan fungsinya.
d. Kelas IV : Air yang dipakai untuk mengairi tanaman dan kebutuhan lain yang tetap harus sesuai dengan standar mutu untuk penggunaannya.

Kualitas air ditentukan berdasarkan jika air tersebut dalam keadaan kondisi normal yaitu kondisi air yang dapat digunakan sesuai fungsi dan peruntukannya (Saputra et al., 2023). Jika air itu menyimpang dari kondisi normal maka dikategorikan air tercemar dan tidak dapat digunakan sesuai fungsi dan peruntukannya. Kondisi tercemar yaitu apabila air tersebut dapat memberi dampak buruk bagi lingkungan sekitar dan apabila kualitas air menyimpang dari standar baku mutu yang seharusnya. Misalkan, jumlah bakteri E. Coli adalah salah satu parameter penting untuk air minum. Mengacu pada Peraturan Menteri Kesehatan Republik Indonesia Nomor 492 Tahun 2010 tentang Persyaratan Kualitas Air Minum, ambang batas maksimum keberadaan bakteri E. Coli ditetapkan sebesar 0 dalam setiap 100 ml sampel air. Apabila hasil analisis menunjukkan adanya satu

koloni bakteri E. Coli dalam volume tersebut, maka air tersebut dinyatakan tidak layak untuk dikonsumsi karena berpotensi membahayakan kesehatan manusia.

2.2.2 Standar Kualitas Air Minum

Kualitas air sangat penting untuk diketahui mengenai komposisi secara kualitatif dan kuantitatif yaitu senyawa kimia, faktor fisika, maupun mikroorganisme yang terkandung di dalam air tersebut (Saputra et al., 2023). Parameter kualitas air berbeda-beda pada tiap jenis airnya, dimana parameter kualitas air merupakan indikator yang menyatakan standar kelayakan air tersebut digunakan sesuai peruntukan dan kebutuhannya. Hal ini sangat perlu diketahui bagi pengguna air, baik secara individu, komunitas, dan industri. Contohnya parameter air untuk kebutuhan industri dan air kolam renang (rekreasi) memiliki parameter yang berbeda dengan air minum. Begitu juga dengan air untuk kebutuhan budidaya pertanian, air limbah industri hingga air limbah domestik, tentu memiliki parameter kualitas yang berbeda-beda satu sama lain.

Parameter wajib sebagai persyaratan kualitas air dibagi menjadi tiga bagian, yaitu:

1. Parameter fisika, yaitu parameter kualitas air yang dapat diamati atau dianalisis berdasarkan karakteristik visual dan fisik. Contohnya suhu, rasa, bau, kekeruhan, warna, dan TSS (*Total Suspended Solid*).
2. Parameter kimia merupakan salah satu aspek yang digunakan untuk menilai kualitas air, yang mencakup kandungan berbagai zat kimia di dalamnya. Unsur-unsur ini bisa berupa senyawa organik maupun anorganik. Beberapa contoh dari parameter ini antara lain nilai pH, tingkat kesadahan, alkalinitas,

kadar oksigen terlarut, TDS (Total Zat Padat Terlarut), BOD (Kebutuhan Oksigen Biologis), COD (Kebutuhan Oksigen Kimia), serta keberadaan logam berat seperti besi, mangan, timbal, raksa, krom, tembaga, dan senyawa kimia lainnya yang dapat mempengaruhi mutu air.

3. Parameter biologi, yaitu parameter kualitas air yang di dalamnya terkandung mikroorganisme. Contohnya yaitu berupa jamur, virus, bakteri pathogen dan ganggang (Saputra et al., 2023).

Agar bisa langsung diminum, air minum harus memenuhi standar kesehatan tertentu, baik itu melalui proses pengolahan maupun tanpa pengolahan. Standar ini mencakup berbagai aspek seperti kondisi fisik, kandungan kimia, unsur biologis, hingga zat radioaktif yang mungkin ada di dalam air. Tujuannya adalah untuk mencegah penyebaran penyakit yang bisa muncul akibat air yang sudah tercemar bahan kimia atau mikroorganisme berbahaya. Karena itu, masyarakat perlu menggunakan air yang sesuai dengan standar kualitas air minum yang telah ditetapkan oleh Kementerian Kesehatan melalui Peraturan Menteri Kesehatan No. 492/Menkes/Per/IV/2010. Rincian parameter kualitas air minum tersebut ditampilkan pada Tabel 2.2.

Tabel 2.2 Parameter Kualitas Air Minum

No.	Jenis Parameter	Satuan	Kadar maksimum yang ditentukan
1.	Parameter Fisik		
	a. Bau		Tidak berbau
	b. Warna	TCU	15
	c. Total zat padat terlarut (TDS)	mg/l	500
	d. Kekeruhan	NTU	5
	e. Rasa		Tidak berasa
	f. Suhu	°C	Suhu udara ±3°
2.	Parameter Kimia		
	a. Aluminium	mg/l	0,2
	b. Besi	mg/l	0,3

	c. Kesadahan/*Hardness* (kapasitas air untuk mengendapkan sabun)	mg/l	500
	d. Khlorida	mg/l	250
	e. Mangan	mg/l	0,4
	f. Ph	mg/l	6,5 – 8,5
	g. Seng	mg/l	3
	h. Sulfat	mg/l	250
	i. Tembaga	mg/l	2
	j. Amonia	mg/l	1,5
	k. Trihalomethanes	µg/l	0,8
	l. Kloramin (Disinfektan Klorin)	mg/l	5
	m. Karbon Organik	mg/l	10
3.	Daya hantar listrik (Konduktivitas)	µS/cm	20 - 1500

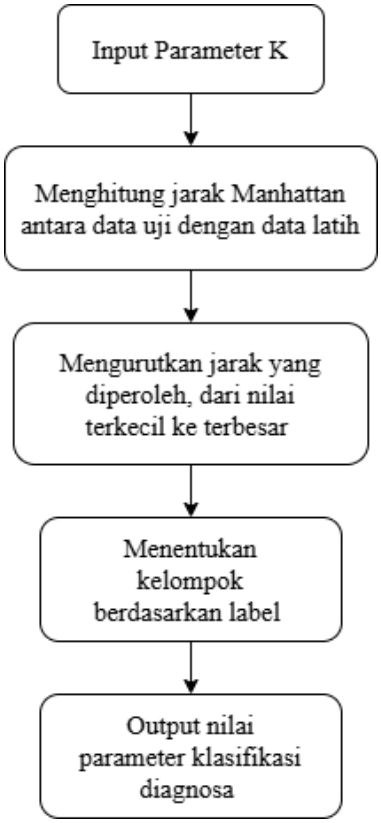
Dengan mengetahui parameter kualitas air, maka dapat dijadikan sebagai acuan dalam melakukan pengendalian kualitas air terutama dalam pengolahan air.

2.3 Algoritma *K-Nearest Neighbor*

Algoritma KNN adalah merupakan salah satu metode klasifikasi yang bekerja dengan menentukan jarak terdekat antara data uji dan data latih. Proses klasifikasi dilakukan dengan menghitung jarak terkecil antara sampel baru terhadap data pelatihan untuk mengidentifikasi tetangga terdekatnya (Swantika et al., 2020). Algoritma KNN dapat memprediksi kasus baru berdasarkan ukuran kesamaan (Sarker et al., 2020). Tujuan utama KNN adalah mengklasifikasikan objek baru berdasarkan atribut dan sampel data yang telah tersedia. Hasil sampel uji baru diklasifikasikan menurut sebagian besar kategori pada algoritma KNN. Dalam proses klasifikasi, KNN hanya didasarkan pada memori dan tidak memakai model apapun untuk disesuaikan. Algoritma KNN memprediksi sampel data uji baru menggunakan klasifikasi ketetanggaan.

Manhattan distance dapat diperoleh dengan metode KNN dengan mengurutkan jarak tertangga terdekat untuk mendapatkan model. Alur proses pada

metode *K-Nearest Neighbor* (KNN) dapat dilihat pada Gambar 2.1 Alur algoritma KNN.



Gambar 2.1 Alur algoritma KNN

Selanjutnya, proses klasifikasi menggunakan metode *K-Nearest Neighbor* (KNN) dilakukan dengan menghitung jarak Manhattan antara data uji dan data latih. Tahapan umum dalam penerapan algoritma KNN untuk klasifikasi adalah sebagai berikut:

1. Input data latih dan data uji
2. Menentukan nilai k
3. Menghitung jarak ketetanggan data uji dan data latih. Rumus perhitungan jarak *Manhattan* dapat dilihat pada persamaan 2.1 (Pangestu & Fitriani, 2022).

$$d_{(x,y)} = \sum_{i=1}^{n} |x_i - y_i| \quad (2.1)$$

Keterangan:

$d_{(x,y)}$: Jarak Manhattan antara data latih dan data uji
$|x_i - y_i|$: nilai absolut dari selisih antara fitur ke-*i*
x_i : atribut dari data ke-*i*, (*i* = 1, 2, 3, …, …., *n*)
y_i : atribut pusat *Cluster* ke-*i*, (*i* = 1, 2, 3, …, …., *n*)

4. Data dikelompokkan berdasarkan hasil dari *Manhattan Distance*.
5. Data dikelompokkan berdasarkan nilai k.
6. Memilih nilai yang paling banyak muncul dari tetangga terdekat sebagai acuan dalam memprediksi data berikutnya.

Berikut adalah *pseudocode* algoritma KNN dengan perhitungan jarak Manhattan sebagai berikut.

```
Input:
1. Data training: X_train (fitur), y_train (label)
2. Data uji: X_test
3. Jumlah tetangga terdekat: k
Output:
1.1.1  Prediksi label untuk setiap data di X_test

Function KNN_Manhattan(X_train, y_train, X_test, k):
  hasil_prediksi = []

  // Loop data uji
  For each x_uji in X_test:
    daftar_jarak = []

    // Looping semua data latih untuk menghitung jaraknya ke
x_uji
    For each i in 0 to length(X_train) - 1:
      x_latih = X_train[i]

      // Perhitungan Manhattan Distance antara x_uji dan x_latih
      jarak = 0
      // Looping tiap fitur (kolom) pada x_uji dan x_latih
      For j in 0 to length(x_uji) - 1:
        //menjumlahkan nilai absolut selisihnya
        jarak += abs(x_uji[j] - x_latih[j])
      EndFor

      // Simpan jarak dan label dari data latih tersebut
      daftar_jarak.append((jarak, y_train[i]))
```

```
    EndFor

    // Urutkan berdasarkan jarak terdekat
    daftar_jarak.sort by jarak ascending

    // Ambil k data dengan jarak terdekat
    k_tetangga = daftar_jarak[0:k]

    // Hitung jumlah label dari k tetangga
    hitung_label = dictionary kosong
    //Looping setiap label dari k_tetangga
    For each (jarak, label) in k_tetangga:
      // hitung berapa kali tiap label muncul
      if label not in hitung_label:
        hitung_label[label] = 1
      else:
        hitung_label[label] += 1
    EndFor

    // Ambil label dengan jumlah terbanyak (mayoritas)
    label_terpilih = label dengan frekuensi tertinggi di
hitung_label

   // Menambahkan hasil prediksi dari x_uji ke dalam list
hasil_prediksi
    hasil_prediksi.append(label_terpilih)
  EndFor

  Return hasil_prediksi
EndFunction
```

Berdasarkan rumus *Manhattan Distance* pada persamaan 2.1, jarak Manhattan antara data latih dan data uji atau $d_{(x,y)}$ diinisialisasi dengan variabel "jarak" pada *pseudocode* algoritma *K-Nearest Neighbor*. Sedangkan x_i (atribut dari data ke-*i*), diinisialisasikan dengan "x_uji[j]" dan y_i (atribut pusat *Cluster* ke-*i*) diinisialisasikan dengan "x_latih[j]".

2.4 Evaluasi Model

Evaluasi model dilakukan untuk memahami kinerja algoritma KNN dalam menentukan akurasi model maka diperlukan *confusion matrix* yang dapat

digunakan khususnya pada kasus klasifikasi dalam mengukur kinerja suatu model pada *machine learning* atau yang disebut dengan *supervised learning* (Said et al., 2022). Berdasarkan hasil klasifikasi sebenarnya, *Confusion matrix* digunakan untuk melihat perbandingan antara hasil klasifikasi yang dilakukan oleh sistem dengan kondisi sebenarnya. Dengan menghitung nilai akurasi dari *confusion matrix*, kita bisa mengetahui seberapa baik performa model yang telah dibuat dalam memprediksi data. Jadi, alat ini membantu kita menilai apakah model sudah bekerja dengan cukup baik atau masih perlu diperbaiki. (Tangkelayuk & Evangs, 2022). *Confusion Matrix* digunakan untuk menghitung *accuracy*. *Confusion Matrix* menghasilkan 4 hasil pengukuran yang terdapat nilai actual dan nilai prediksi. Tabel pada Evaluasi Model ini ditunjukkan pada pada Tabel 2.3.

Tabel 2.3 Tabel Confusion Matrix untuk Evaluasi Model

Confusion Matrics		Nilai Aktual	
		Positif	**Negatif**
Nilai Prediksi	**Positif**	True Positive	False Positive
	Negatif	False Negative	True Negative

Keterangan :

a. *True Positives* (TP): Data yang seharusnya masuk kategori positif dan berhasil diprediksi dengan tepat oleh sistem.
b. *False Positives* (FP): Data yang sebenarnya negatif, tapi sistem salah memprediksinya sebagai positif.
c. *False Negatives* (FN): Data yang sebenarnya positif, tapi malah diprediksi sebagai negatif oleh sistem.

d. *True Negatives* (TN): Data yang memang negatif dan berhasil diprediksi dengan benar oleh sistem.

Untuk mengetahui seberapa baik kinerja algoritma KNN dalam melakukan prediksi, digunakan confusion matrix sebagai alat evaluasi. Confusion matrix ini menunjukkan perbandingan antara hasil prediksi sistem dan kondisi sebenarnya dari data. Dari situ, bisa dihitung beberapa ukuran penting seperti akurasi, presisi, *recall*, dan *F1-score* yang digunakan untuk menilai ketepatan model. Rumus dari masing-masing pengukuran tersebut dapat dilihat pada Tabel 2.4.

Tabel 2.4 Pengukuran Nilai Akurasi, Presisi, *Recall*, dan *F1-Score*

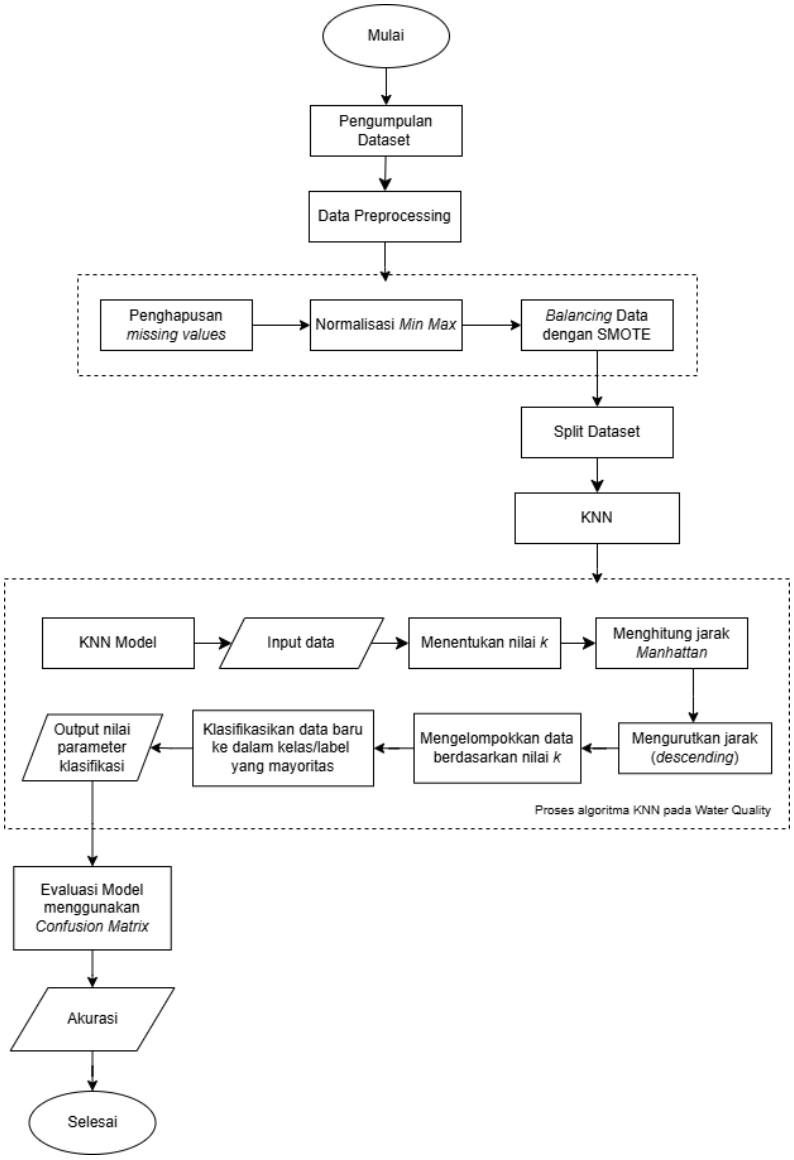
No.	**Pengukuran**	**Rumus**
1.	Akurasi	$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
2.	Presisi	$\text{Precision} = \frac{TP}{TP+FP}$
3.	*Recall*	$Recall = \frac{TP}{TP+FP}$
4.	*F1-Score*	$F1\text{-}Score = \frac{2\ x\ Recall\ x\ Precision}{Recall+Precision}$

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Sistem

Penelitian ini menggunakan metode KNN untuk mengelompokkan data berdasarkan kualitas airnya. Dengan metode ini, sistem bisa menentukan apakah air tergolong layak atau tidak untuk dikonsumsi, berdasarkan kemiripan data baru dengan data yang sudah ada sebelumnya. Desain sistem yang diterapkan dalam riset ini ditunjukkan melalui alur diagram di Gambar 3.1.



Gambar 3.1 Desain Sistem

Desain sistem pada Gambar 3.1 diawali dengan penginputan data kualitas air, kemudian melakukan *preprocessing* yaitu menghapus data-data yang terdapat nilai kosong, proses normalisasi untuk menghasilkan data dengan nilai rentang antara 0 dan 1 menggunakan *Min Max Scaling*, lalu menyeimbangkan kelas data dengan *SMOTE*. Langkah selanjutnya adalah membagi data menjadi dua bagian, yaitu data latih dan data uji. Setelah pembagian selesai, proses klasifikasi dilakukan

menggunakan algoritma KNN. Tahapan awal dalam proses ini dimulai dengan memasukkan data latih ke dalam sistem untuk membentuk model yang akan digunakan dalam pengujian, kemudian menentukan nilai *k*, menghitung *Manhattan Distance*, mengurutkan jarak secara *descending*, mengelompokkan data berdasarkan nilai *k*, mengelompokkan data ke dalam label/kelas yang mayoritas lalu menghasilkan output nilai parameter klasifikasi. Langkah terakhir dilanjutkan dengan evaluasi model dengan *Confusion Matrix* yang akan menghasilkan akurasi.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini diambil dari situs www.kaggle.com, dan pertama kali diunggah oleh Aditya Kadiwal pada tahun 2021 dengan judul *Water Potability*. Dataset tersebut berformat csv dan berisi 10 variabel, yang terdiri dari 9 atribut sebagai input dan 1 label sebagai penentu kelas. Rincian mengenai parameter-parameter pada data kualitas air ini dapat dilihat pada Tabel 3.1.

Tabel 3.1 Parameter Data Kualitas Air

No.	Parameter Kualitas Air									
	Ph	**Hardness**	**Solids**	**Chloramines**	**Sulfate**	**Conductivity**	**Organic Carbon**	**Trihalomethanes**	**Turbidity**	**Potability**
1.		204.89	20791.32	7.30	368.52	564.31	10.38	86.99	2.96	0
2.	3.72	129.42	18630.06	6.64		592.89	15.18	56.33	4.50	0
3.	8.10	224.24	19909.54	9.28		418.61	16.87	66.42	3.06	0
4.	8.32	214.37	22018.42	8.06	356.89	363.27	18.44	100.34	4.63	0
5.	9.09	181.10	17978.99	6.55	310.14	398.41	11.56	32.00	4.08	0
6.	5.58	188.31	28748.69	7.54	326.68	280.47	8.40	54.92	2.56	0
7.	10.22	248.07	28749.72	7.51	393.66	283.65	13.79	84.60	2.67	0
8.	8.64	203.36	13672.09	4.56	303.31	474.61	12.36	62.80	4.40	0
9.		118.99	14285.58	7.80	268.65	389.38	12.71	53.93	3.60	0
10.	11.18	227.23	25484.51	9.08	404.04	563.89	17.93	71.98	4.37	0
11.	7.36	165.52	32452.61	7.55	326.62	425.38	15.59	78.74	3.66	0
...	...	...	...	...	...	...	...	...	...	...
3271.	6.07	186.66	26138.78	7.75	345.70	415.89	12.07	60.42	3.67	1
3272.	4.67	193.68	47580.99	7.17	359.95	526.42	13.89	66.69	4.44	1
3273.	7.81	193.55	17329.8	8.06		392.45	19.90		2.80	1
3274.	9.42	175.76	33155.58	7.35		432.04	11.04	69.85	3.30	1
3275.	5.13	230.60	11983.87	6.30		402.88	11.17	77.49	4.71	1
3276.	7.87	195.10	17404.18	7.51		327.46	16.14	78.70	2.31	1

Berdasarkan Tabel 3.1, besar jumlah data yang digunakan yaitu 3.276 baris data dengan sepuluh atribut dan beberapa parameter pada data kualitas air diantaranya:

1. *ph*: Tingkat keasaman air (0 sampai 14) dalam satuan *mg/L*
2. Kekerasan (*Hardness*): Menunjukkan seberapa besar kemampuan air dalam membentuk endapan sabun, biasanya diukur dalam mg/L.
3. TDS (*Solids*): Jumlah zat padat yang terlarut dalam air, diukur dalam satuan ppm.
4. Kloramin (*Chloramines*): Kandungan zat kloramin dalam air, dinyatakan dalam ppm.
5. Sulfat (*Sulfate*): Seberapa banyak zat sulfat yang larut dalam air, juga dalam ppm.
6. Konduktivitas (*Conductivity*): Menggambarkan kemampuan air dalam menghantarkan listrik, dinyatakan dalam µS/cm.
7. Karbon Organik (*Organic Carbon*): umlah karbon dari bahan organik yang ada dalam air, satuannya ppm.
8. *Trihalomethanes* (THMs): Kandungan senyawa kimia THM dalam air, biasanya dinyatakan dalam µg/L.
9. Kekeruhan air *(Turbidity*): Mengukur tingkat kejernihan atau kekeruhan air, satuannya NTU.
10. *Potability* (Kelayakan konsumsi): Menunjukkan apakah air tersebut aman untuk diminum atau tidak. (0 = tidak layak diminum, 1 = layak diminum).

Pada Tabel 3.1 Parameter Data Kualitas Air, Atribut ph sampai dengan kekeruhan merupakan tipe data *real* dan atribut potabilitas merupakan tipe data *integer*. Satuan *ppm* pada beberapa parameter data kualitas air merupakan singkatan dari *Parts per Million* atau bagian per sejuta, dimana satu *parts per million* sama dengan 0,0001 persen larutan.

Data kualitas air dalam penelitian ini dapat diketahui melalui *Exploratory Data Analysis* (EDA) untuk memahami struktur, distribusi, dan pola data sebelum dilakukan *modeling*. Untuk mengetahui kondisi data tersebut, dilakukan beberapa analisis EDA yaitu sebagai berikut:

3.2.1 Visualisasi Missing Value pada Data Kualitas Air

Data-data yang terdapat nilai yang kosong pada dataset kualitas air ditunjukkan pada Tabel 3.2 sebagai berikut.

Tabel 3.2 Data kualitas air

No.	Parameter Kualitas Air									
	Ph	**Hardness**	**Solids**	**Chloramines**	**Sulfate**	**Conductivity**	**Organic Carbon**	**Trihalomethanes**	**Turbidity**	**Potability**
1.	NaN	204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
2.	3.71608	129.4229	18630.06	6.635246	NaN	592.8854	15.18001	56.32908	4.500656	0
3.	8.099124	224.2363	19909.54	9.275884	NaN	418.6062	16.86864	66.42009	3.055934	0
4.	8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
5.	9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
6.	5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
...	...	...	...	...	...	...	...	...	...	...
3272.	4.668102	193.6817	47580.99	7.166639	359.9486	526.4242	13.89442	66.68769	4.435821	1
3273.	7.808856	193.5532	17329.8	8.061362	NaN	392.4496	19.90323	NaN	2.798243	1
3274.	9.41951	175.7626	33155.58	7.350233	NaN	432.0448	11.03907	69.8454	3.298875	1
3275.	5.126763	230.6038	11983.87	6.303357	NaN	402.8831	11.16895	77.48821	4.708658	1
3276.	7.874671	195.1023	17404.18	7.509306	NaN	327.4598	16.14037	78.69845	2.309149	1

Berdasarkan dataset kualitas air pada Tabel 3.2, beberapa data terdapat *missing value* atau data dengan nilai yang kosong cukup banyak yang dinisialisasikan dengan nilai = (NaN) atau tidak diketahui. Tersebar pada beberapa fitur atau atribut yaitu pada atribut ph, *sulfate*, dan *trihalomethanes*. Berikut akan ditampilkan *Missing Value Map* (peta nilai hilang) pada dataset ditunjukkan di Gambar 3.2.

Gambar 3.2 *Missing Value Map*

Gambar 3.2 tersebut yaitu Visualisasi sebaran data yang hilang di dalam dataset. Berdasarkan visualisasi *Missing Value Map*, terlihat bahwa fitur ph dan *Sulfate* memiliki jumlah nilai hilang yang cukup signifikan (sekitar 20%), sedangkan *Trihalomethanes* dan Karbon Organik memiliki proporsi nilai hilang yang lebih rendah. Berikut ini ditampilkan jumlah data yang kosong pada dataset ditunjukkan pada Tabel 3.3.

Tabel 3. 3 Jumlah data yang kosong pada Data Kualitas Air

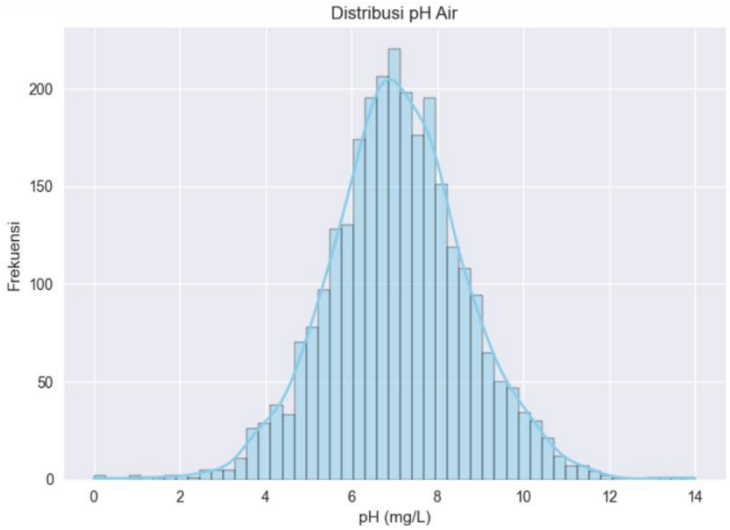
No.	Parameter	Jumlah data yang kosong
1.	Ph	491
2.	Hardness	0
3.	Solids	0
4.	Chloramines	0
5.	Sulfate	781
6.	Conductivity	0
7.	Organic Carbon	0
8.	Trihalomethanes	162
9.	Turbidity	0
10.	Potability	0

Jumlah data yang terdapat nilai yang kosong ditunjukkan pada Tabel 3.3. Data yang kosong terdapat pada 3 parameter yaitu parameter Ph berjumlah 491 data, parameter *Sulfate* berjumlah 781 data, dan parameter *Trihalomethanes* berjumlah 162 data.

3.2.2 Grafik Distribusi

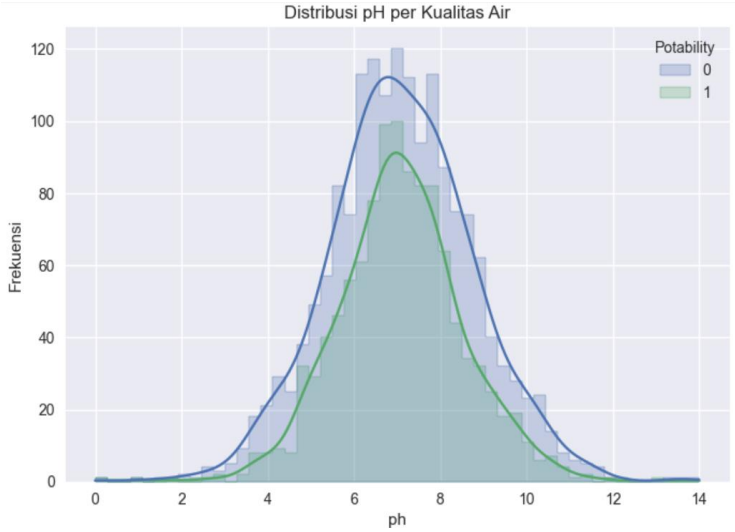
3.2.2.1 Grafik Distribusi pH Air

Berikut grafik distribusi pada parameter pH data kualitas air ditunjukkan pada Gambar 3.3.



Gambar 3. 3 Distribusi Ph Air

Berdasarkan histogram pada Gambar 3.3, nilai pH air tersebar dalam rentang yang cukup luas, yaitu dari sekitar 0 hingga 14. Namun, sebagian besar data terkonsentrasi di kisaran 6,5 hingga 8, yang merupakan rentang pH netral. Distribusi ini menyerupai distribusi normal, dengan puncak frekuensi di sekitar pH 7. Hal ini mengindikasikan bahwa sebagian besar sampel air memiliki tingkat keasaman atau Tingkat basa yang wajar dan sesuai dengan batas standar air minum.

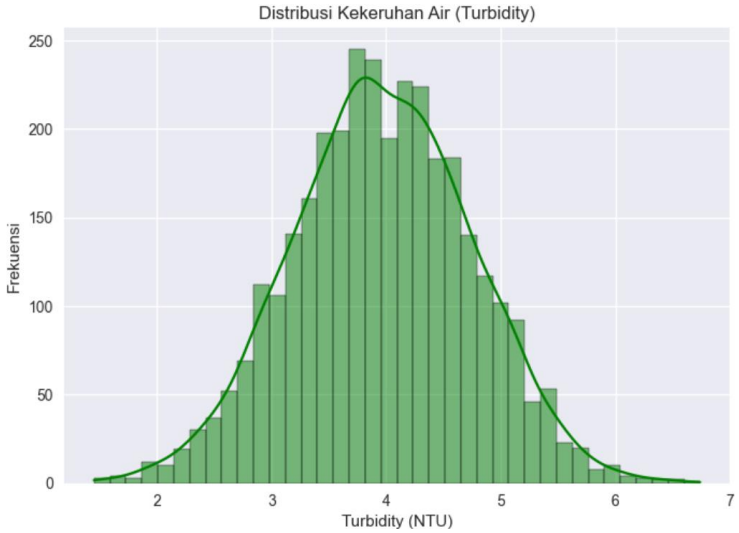


Gambar 3. 4 Distribusi Ph per Kualitas Air

Gambar 3.4 memperlihatkan distribusi pH berdasarkan label kualitas air (*potability*). Ph menunjukkan perbedaan yang paling signifikan antara air layak dan tidak layak minum. pH yang berada dalam kisaran netral (6,5–8) lebih sering diasosiasikan dengan air yang layak konsumsi, sedangkan nilai ekstrem cenderung berasosiasi dengan air yang tidak layak. Fitur ini memiliki distribusi yang relatif normal, sehingga sangat sesuai untuk digunakan dalam model klasifikasi seperti *K-Nearest Neighbor* (KNN).

3.2.2.2 Grafik Distribusi Kekeruhan Air (*Turbidity*)

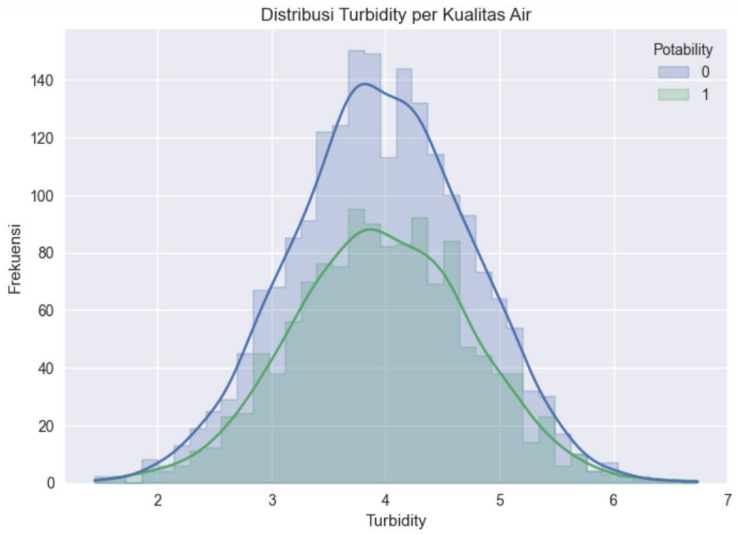
Berikut grafik distribusi pada parameter Kekeruhan Air (*Turbidity*) data kualitas air ditunjukkan pada Gambar 3.5.



Gambar 3.5 Distribusi Kekeruhan Air (*Turbidity*)

Berdasarkan gambar 3.5 distribusi *turbidity* air menunjukkan pola yang hampir menyerupai normal, dengan mayoritas data berada pada kisaran 3 hingga 5 NTU (*Nephelometric Turbidity Unit*). Sebagian besar air memiliki tingkat

kekeruhan yang relatif rendah hingga sedang. Berikut grafik distribusi Kekeruhan Air (*Turbidity*) berdasarkan kelas *potability* ditunjukkan pada Gambar 3.6.

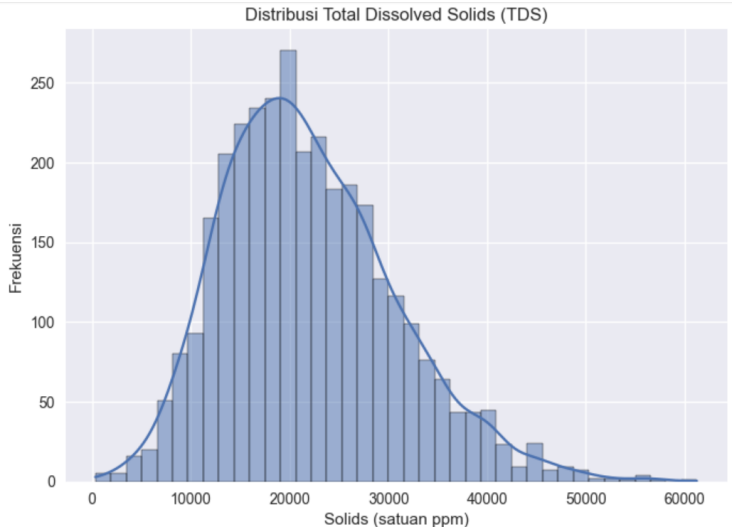


Gambar 3. 6 Distribusi Kekeruhan Air (*Turbidity*) per Kualitas Air

Gambar 3.6 menunjukkan perbedaan distribusi *turbidity* antara air layak dan tidak layak minum. Air layak minum cenderung memiliki *turbidity* yang lebih rendah dan lebih terpusat, sedangkan air tidak layak minum memiliki distribusi yang lebih menyebar dan mencakup nilai *turbidity* yang lebih tinggi. Hal ini tetap menunjukkan bahwa tingkat kekeruhan dapat menjadi salah satu indikator kualitas air yang baik.

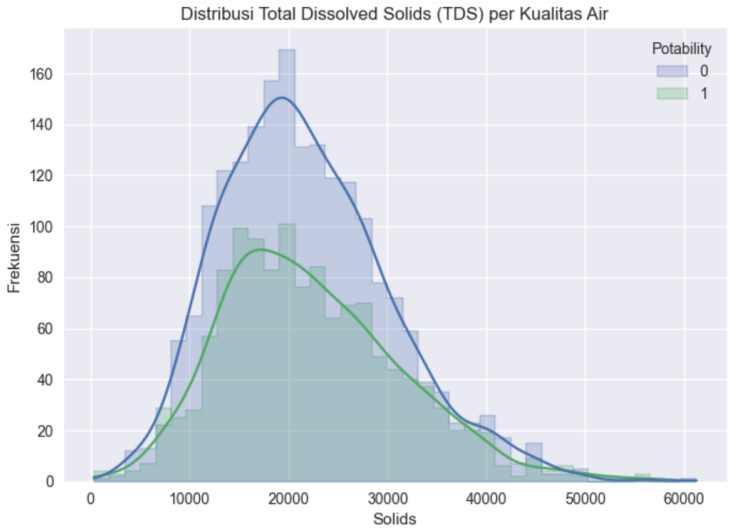
3.2.2.3 Grafik Distribusi *Total Dissolved Solids* (TDS)

Berikut grafik distribusi pada Total padatan terlarut dalam satuan *ppm* data kualitas air ditunjukkan pada Gambar 3.7.



Gambar 3. 7 Distribusi Total padatan terlarut dalam satuan ppm

Gambar 3.7 menunjukkan sebagian besar nilai TDS terkonsentrasi pada kisaran 10.000 hingga 25.000 ppm, dengan puncak sekitar 20.000 ppm. Terdapat *outlier* atau nilai ekstrem yang mencapai lebih dari 60.000 ppm, meskipun frekuensinya sangat rendah. Distribusi ini mengindikasikan bahwa sebagian besar sampel air memiliki tingkat kelarutan zat padat yang masih berada dalam rentang umum, namun ada variasi yang cukup besar antar sampel. Berikut grafik distribusi Total padatan terlarut berdasarkan nilai *potability* ditunjukkan pada Gambar 3.8.

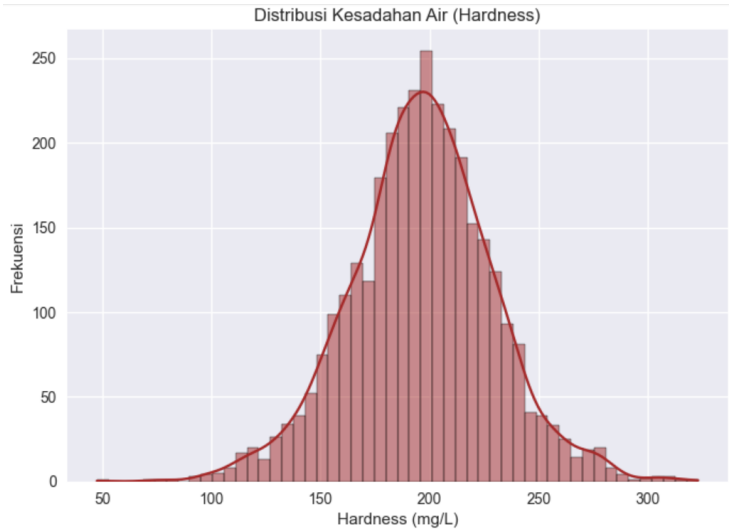


Gambar 3.8 Distribusi TDS per Kualitas Air

Gambar 3.8 menunjukkan Histogram TDS berdasarkan nilai *Potability* (0 = tidak layak, 1 = layak). Air tidak layak konsumsi (*Potability* = 0) cenderung memiliki sebaran nilai TDS yang lebih luas dan puncak yang lebih tinggi di kisaran 19.000–21.000 ppm. Sementara air layak konsumsi (*Potability* = 1) memiliki distribusi yang lebih merata dan landai, dengan puncak sedikit lebih rendah dan tersebar di sekitar 15.000-20.000 ppm. Hal ini mengindikasikan bahwa TDS yang terlalu tinggi berasosiasi dengan air yang tidak layak konsumsi, namun nilai TDS yang sangat rendah juga tidak menjamin kelayakan konsumsi.

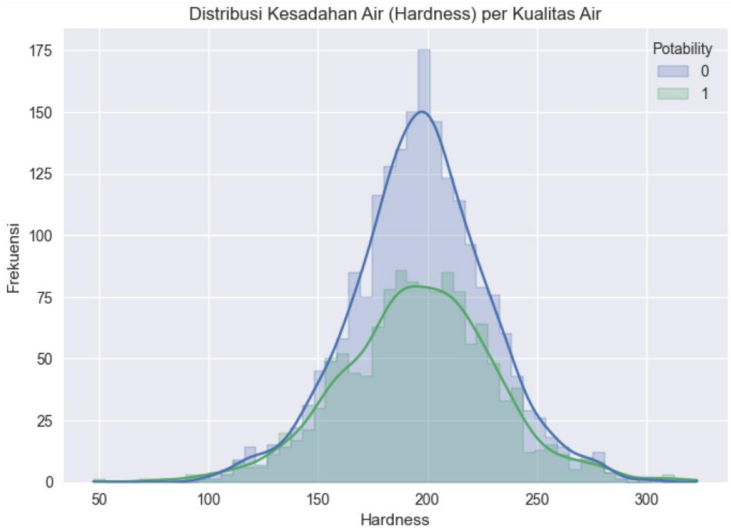
3.2.2.4 Grafik Distribusi Kesadahan Air (*Hardness*)

Berikut grafik distribusi pada kesadahan air dalam satuan mg/L pada data kualitas air ditunjukkan pada Gambar 3.9.



Gambar 3.9 Distribusi Kesadahan Air

Distribusi nilai kesadahan air menunjukkan pola yang lebih simetris dan mendekati distribusi normal. Terlihat pada puncak distribusi terjadi pada nilai sekitar 200 mg/L. Gambar 3.9 menunjukkan bahwa nilai-nilai *Hardness* tersebar relatif merata di kedua sisi rata-rata. Berikut grafik Distribusi *Hardness* Berdasarkan Kualitas Air ditunjukkan pada Gambar 3.10.

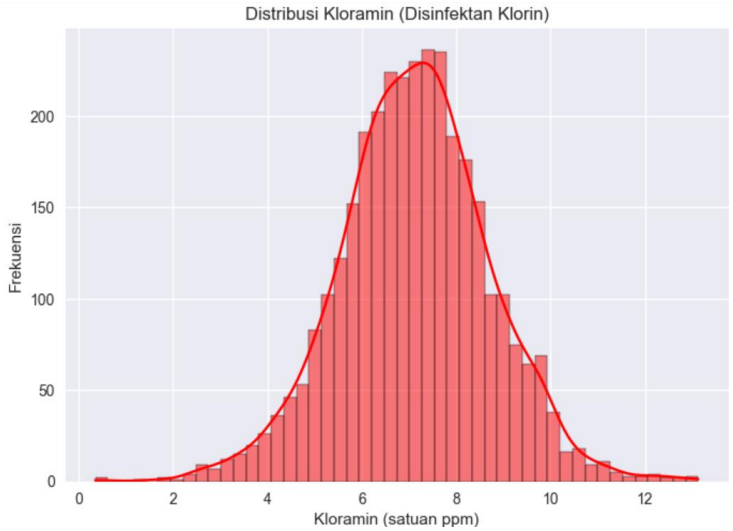


Gambar 3.10 Distribusi *Hardness* per Kualitas Air

Berdasarkan grafik pada gambar 3.10, air yang tidak layak konsumsi (*Potability* = 0) memiliki puncak distribusi sedikit lebih tinggi, dengan konsentrasi pada rentang 190–210 mg/L. Sedangkan, air yang layak konsumsi (*Potability* = 1) menunjukkan puncak yang sedikit lebih rendah dan lebar distribusi yang lebih datar. Parameter *Hardness* menunjukkan distribusi yang hampir normal dan tidak menunjukkan perbedaan tajam terhadap kualitas air, meskipun tetap dapat menjadi faktor pendukung dalam model klasifikasi.

3.2.2.5 Grafik Distribusi Kloramin (Disinfektan Klorin)

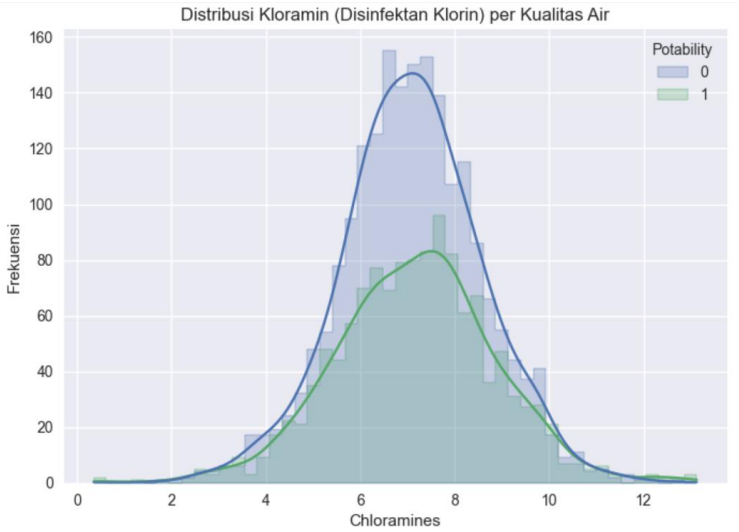
Berikut grafik distribusi pada Disinfektan Klorin dalam satuan ppm pada data kualitas air ditunjukkan pada Gambar 3.11.



Gambar 3.11 Distribusi Kloramin

Chloramines (senyawa berbasis klorin) digunakan sebagai disinfektan. Grafik pada gambar 3.11 ini menunjukkan distribusi *Chloramines* dalam satuan ppm. Terlihat sebaran yang mendekati normal, dengan puncak frekuensi pada rentang 6–8 ppm, dan rentang nilai dari 0 hingga >12 ppm. Ini menunjukkan bahwa

sebagian besar sampel air memiliki kadar kloramin dalam rentang sedang. Berikut grafik Distribusi Kloramin berdasarkan nilai Kualitas Air (*Potability*) ditunjukkan pada Gambar 3.12.

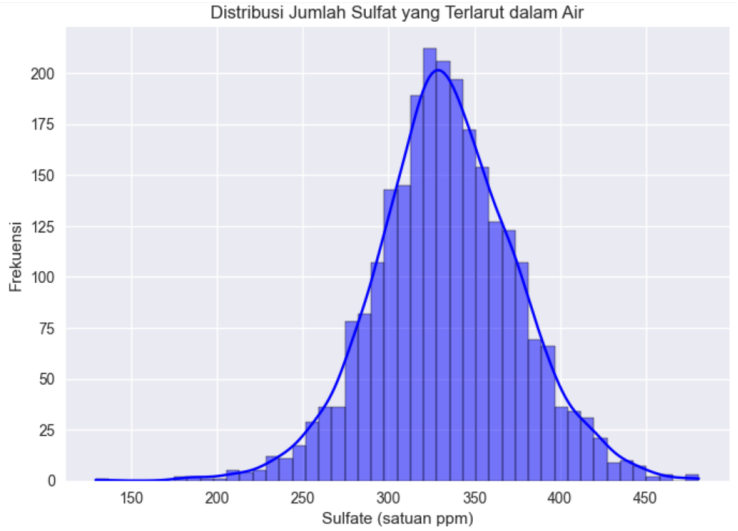


Gambar 3.12 Distribusi Kloramin per Kualitas Air

Grafik pada gambar 3.12 menunjukkan Air tidak layak minum (biru) cenderung memiliki puncak kloramin lebih sempit dan tinggi pada sekitar 7 ppm. Sedangkan air layak minum (hijau) memiliki distribusi yang lebih menyebar dan puncak lebih rendah, namun tetap berada di sekitar rentang 6–8 ppm. Terlihat bahwa air tidak layak memiliki konsentrasi kloramin yang lebih terkonsentrasi. Variabel *Chloramines* juga mengikuti distribusi normal tetapi menunjukkan perbedaan yang lebih jelas antara air layak dan tidak layak, sehingga memiliki potensi sebagai fitur yang lebih informatif dalam model klasifikasi kualitas air.

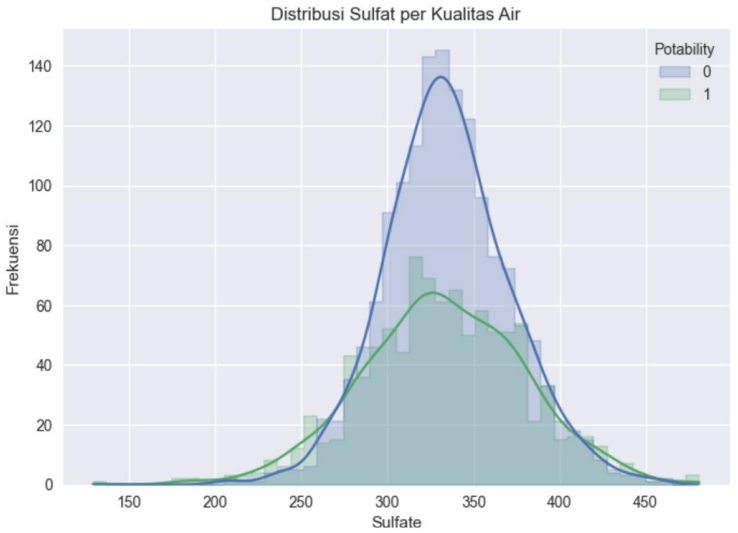
3.2.2.6 Grafik Distribusi Sulfat dalam Air

Berikut grafik distribusi pada jumlah sulfat yang terlarut dalam air ditunjukkan pada Gambar 3.13.



Gambar 3. 13 Distribusi Sulfat yang Terlarut Dalam Air

Grafik pada gambar 3.13 menunjukkan sebaran nilai Sulfat (dalam ppm) dalam seluruh sampel air. Distribusi tampak mendekati normal, dengan puncak di sekitar 330 ppm. Sebaran data cukup simetris, dengan sebagian besar nilai berada di kisaran 250–410 ppm. Tidak terdapat nilai ekstrem (*outlier*) yang jelas, menandakan stabilitas dalam sebaran data sulfat. Berikut grafik Distribusi Sulfat berdasarkan nilai Kualitas Air (*Potability*) ditunjukkan pada Gambar 3.14.

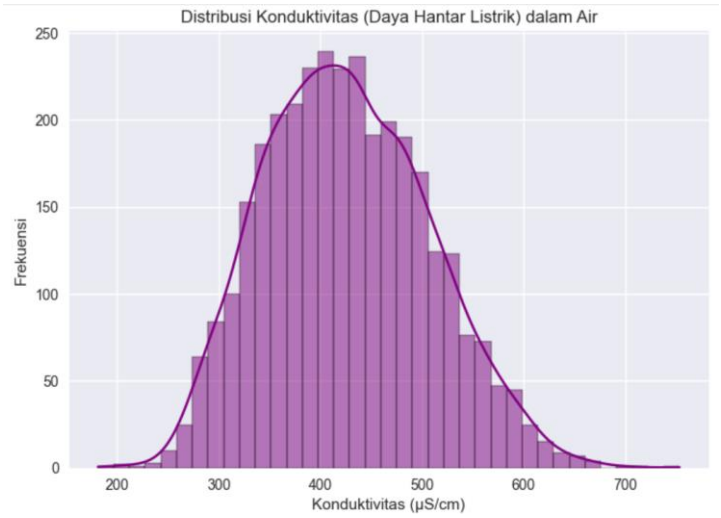


Gambar 3. 14 Distribusi Sulfat per Kualitas Air

Gambar 3.14 menunjukkan Distribusi Sulfat berdasarkan Potabilitas. Air tidak layak minum (label 0) cenderung memiliki nilai sulfat yang lebih tinggi, dengan puncak distribusi mendekati 340 ppm. Air layak minum (label 1) memiliki puncak distribusi di bawahnya, sekitar 320 ppm, dengan penyebaran yang lebih merata dan frekuensi lebih rendah. Kandungan sulfat yang tinggi memungkinkan air tersebut tidak layak minum. Namun, parameter sulfat penting dalam menentukan kualitas air minum, sehingga dapat digunakan sebagai indikator utama dalam model klasifikasi kualitas air minum.

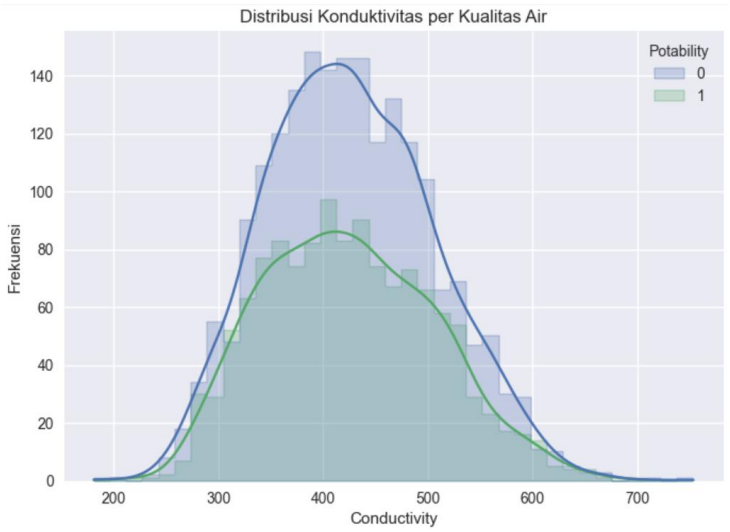
3.2.2.7 Grafik Distribusi Konduktivitas Air

Berikut grafik distribusi pada konduktivitas air ditunjukkan pada Gambar 3.15.



Gambar 3.15 Distribusi Konduktivitas Air

Grafik pada gambar 3.15 menunjukkan histogram distribusi nilai konduktivitas (µS/cm) pada sampel air. Terlihat bahwa sebaran data mendekati distribusi normal, dengan nilai konduktivitas yang paling sering muncul berada di sekitar 450 µS/cm. Hal ini menunjukkan bahwa sebagian besar sampel air memiliki daya hantar listrik yang berada pada kisaran menengah, yang mencerminkan adanya kandungan ion terlarut pada tingkat sedang. Berikut grafik distribusi konduktivitas berdasarkan nilai Kualitas Air (*Potability*) ditunjukkan pada Gambar 3.16.



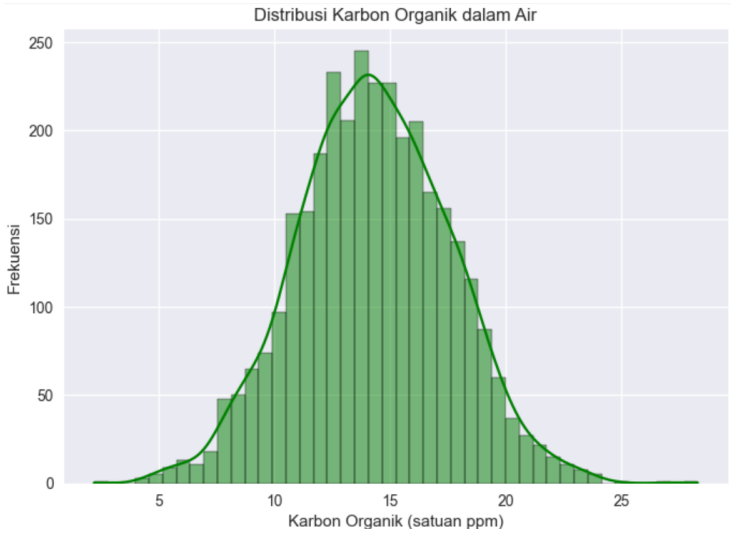
Gambar 3.16 Distribusi Konduktivitas berdasarkan nilai *Potability*

Gambar 3.16 menunjukkan distribusi nilai konduktivitas yang dikelompokkan berdasarkan status potabilitas air. Terlihat bahwa air yang tidak layak minum cenderung memiliki nilai konduktivitas yang lebih tinggi, dengan puncak frekuensi sekitar 470 µS/cm. Sebaliknya, air yang layak minum memiliki distribusi yang lebih merata dan sedikit bergeser ke kiri, dengan puncak sekitar 400 µS/cm. Hal ini menunjukkan adanya hubungan negatif antara nilai konduktivitas dan kelayakan konsumsi air. Semakin tinggi nilai konduktivitas, semakin besar

kemungkinan air tersebut mengandung zat terlarut yang berpotensi membahayakan kesehatan.

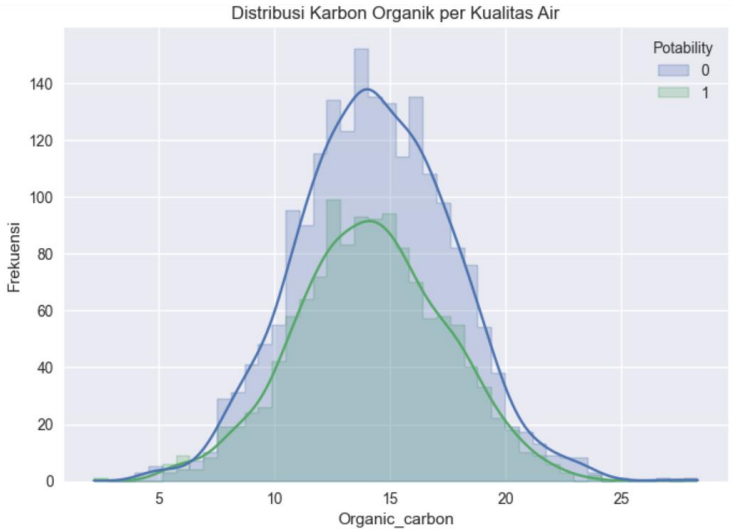
3.2.2.8 Grafik Distribusi Karbon Organik dalam Air

Berikut grafik distribusi pada karbon organik dalam air dalam satuan ppm atau setara dengan mg/L ditunjukkan pada Gambar 3.17.



Gambar 3.17 Distribusi karbon organik dalam air

Distribusi kadar karbon organik dalam air pada gambar 3.17 menunjukkan bentuk yang mendekati distribusi normal dengan puncak sekitar 14–15 ppm. Mayoritas sampel memiliki kadar karbon organik antara 10 hingga 20 ppm. Variasi yang terbentuk masih dalam batas yang wajar, meskipun sedikit lebih condong ke arah kanan yang menunjukkan adanya sejumlah kecil air dengan kadar karbon organik tinggi. Berikut grafik distribusi karbon organik berdasarkan nilai kualitas air (*Potability*) ditunjukkan pada Gambar 3.18.

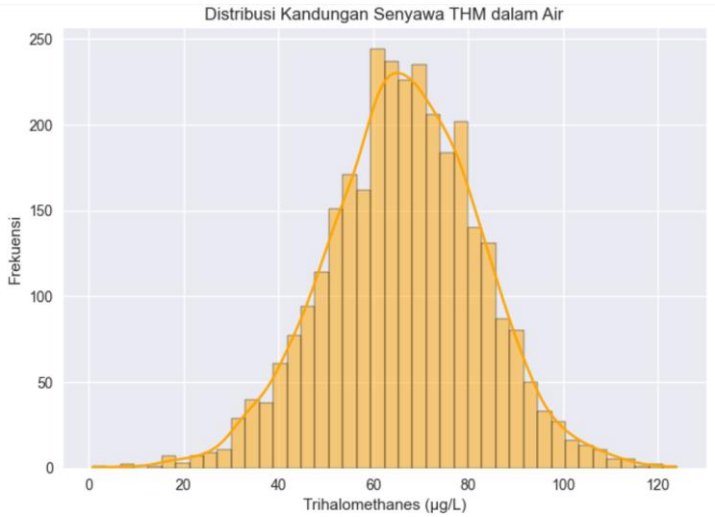


Gambar 3.18 Distribusi karbon organik berdasarkan nilai *Potability*

Hasil distribusi berdasarkan potabilitas air pada gambar 3.18 menunjukkan bahwa air yang tidak layak konsumsi (*potability* = 0) umumnya memiliki kadar karbon organik yang lebih tinggi dibandingkan dengan air yang layak konsumsi (*potability* = 1). Hal ini mengindikasikan adanya hubungan negatif antara kandungan karbon organik dan kualitas air. Semakin tinggi kadar karbon organik, semakin besar kemungkinan air tersebut tidak aman untuk diminum.

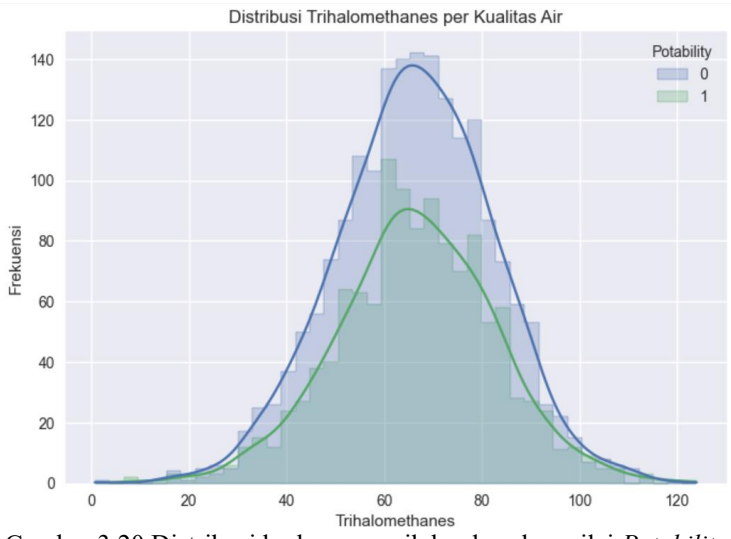
3.2.2.9 Grafik Distribusi Senyawa THM dalam Air

Berikut grafik distribusi Kandungan Senyawa Trihalomethanes dalam air dengan satuan µg/L ditunjukkan pada Gambar 3.19.



Gambar 3.19 Distribusi Senyawa THM dalam Air

Grafik pada gambar 3.19 menunjukkan distribusi frekuensi dari kandungan THM (µg/L) dalam air. Distribusi data terlihat mendekati normal, dengan puncak pada rentang 60–70 µg/L. Terlihat bahwa sebagian besar sampel memiliki kandungan THM antara 40 hingga 90 µg/L. Nilai ekstrim (*outliers*) juga tampak di bagian kiri dan kanan distribusi, namun jumlahnya kecil. Berikut grafik distribusi senyawa Trihalomethanes berdasarkan nilai kualitas air (*Potability*) ditunjukkan pada Gambar 3.20.

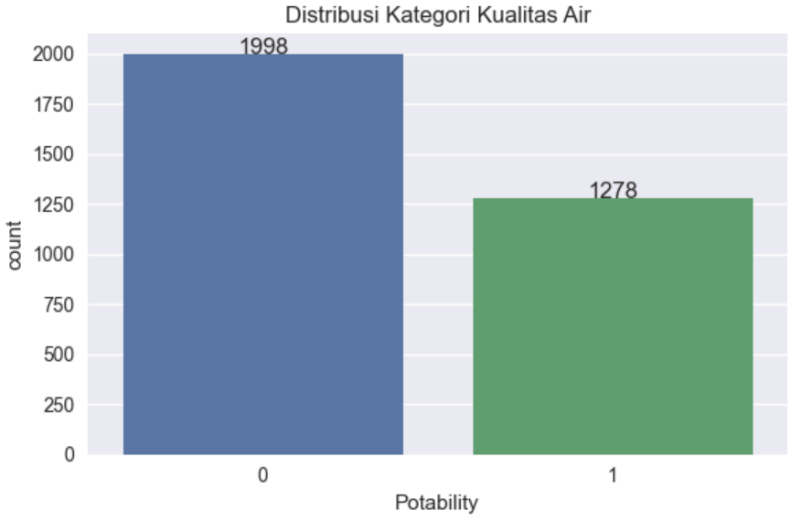


Gambar 3.20 Distribusi karbon organik berdasarkan nilai *Potability*

Pada grafik gambar 3.20, distribusi THM berdasarkan kualitas air. Air tidak layak minum (biru) mengindikasikan kandungan THM yang lebih tinggi secara umum. Sebaliknya, air layak minum (hijau) menunjukkan distribusi THM yang sedikit lebih rendah dan lebih sempit. Hal ini mendukung bahwa tingginya kandungan THM berkaitan dengan ketidaklayakan air untuk diminum.

3.2.2.10 Grafik Distribusi Kategori Kualitas Air (Potabilitas)

Berikut grafik distribusi kategori kualitas air (potabilitas) yang menunjukkan hasil keseimbangan data kualitas air dapat dilihat pada Gambar 3.21.



Gambar 3.21 Distribusi Kategori Kualitas Air (Potabilitas)

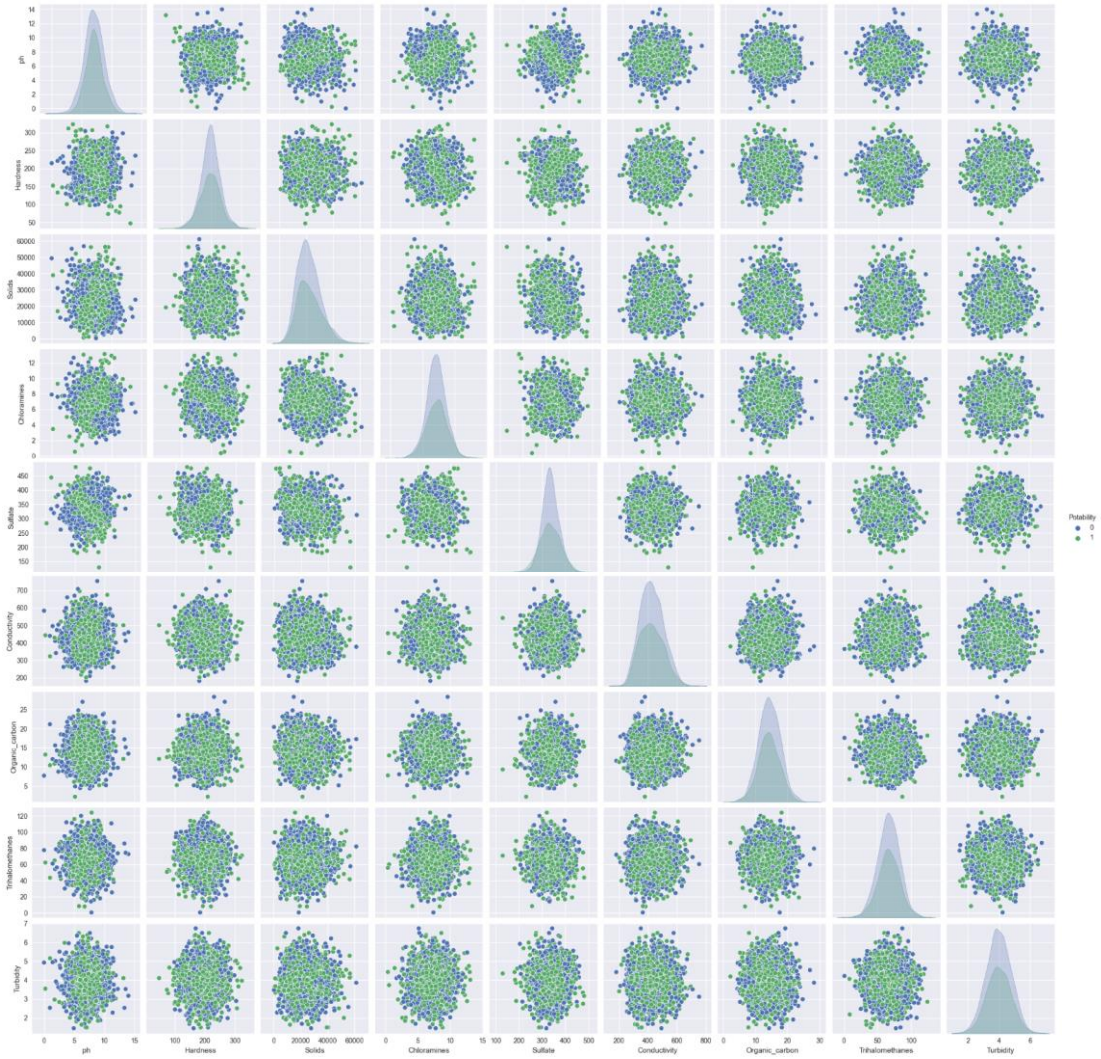
Grafik pada gambar 3.21 menampilkan jumlah sampel air berdasarkan kategorisasi potabilitas yaitu apakah air tersebut layak minum (1) atau tidak layak minum (0). Jumlah sampel air tidak layak minum (*potability* = 0) sebanyak 1.998 sampel ditampilkan dalam warna biru. Jumlah sampel air layak minum (*potability* = 1) sebanyak 1.278 sampel, ditampilkan dalam warna hijau. Terjadi ketidakseimbangan kelas dengan proporsi air tidak layak minum lebih tinggi

dibandingkan air yang layak. Ketidakseimbangan ini perlu diperhatikan jika akan digunakan dalam pemodelan klasifikasi, karena model bisa menjadi bias terhadap kelas mayoritas (air tidak layak minum).

Berdasarkan bahasan 3.2.2.1 Distribusi pH Air hingga 3.2.2.10 Grafik distribusi kategori kualitas air terlihat bahwa grafiknya berhimpitan. Setiap variabel hanya bisa berdiri sendiri karena tidak bisa memisahkan secara langsung. Terlihat pada gambar distribusi dar seluruh variabel berdasarkan nilai kualitas air (*Potability*) menunjukkan kelas air yang layak konsumsi (*potability* = 1) dan kelas air yang tidak layak konsumsi (*potability* = 0) dimana grafik antar kelas berhimpitan, serta kurang efektif dalam membedakan kualitas air karena distribusinya hampir sama pada kedua kelas. Maka dari itu, dalam membuat model klasifikasi KNN tidak bisa hanya mengandalkan salah satu variabel saja karena satu variabel saja tidak cukup untuk menentukan kelayakan air minum, sehingga perlu dipadukan dengan variabel lainnya.

3.2.3 Visualisasi Hubungan antar Variabel

Berikut visualisasi hubungan antar variabel pada data kualitas air dapat dilihat pada Gambar 3.22.



Gambar 3.22 Visualisasi Hubungan antar Variabel

Gambar 3.22 merupakan hasil visualisasi *pairplot* untuk menunjukkan sebaran dan hubungan antar fitur, dengan pewarnaan berdasarkan kelas *Potability*. Warna biru menunjukkan air tidak layak minum (*Potability* = 0) dan warna hijau adalah air layak minum (*Potability* = 1). Visualisasi tersebut menunjukkan bahwa distribusi variabel seperti *Solids*, *Conductivity*, dan *Hardness* tampak hampir normal. Pola penyebaran antar fitur cenderung acak dan tidak membentuk pola jelas. Tidak terlihat *cluster* atau pemisahan visual yang tegas antara data dengan

Potability 0 dan 1, menunjukkan bahwa fitur-fitur tersebut tidak memiliki kekuatan klasifikasi yang tinggi secara individual. Namun, beberapa fitur seperti *Trihalomethanes* dan *Turbidity* menunjukkan sebaran yang sedikit berbeda antara kelas 0 dan 1, meskipun tidak signifikan secara visual.

Maka, dapat disimpulkan visualisasi hubungan antar variabel pada data kualitas air menunjukkan bahwa tidak terdapat fitur yang secara langsung berkorelasi kuat dengan kelas *Potability* disebabkan sebaran data yang acak tanpa pola distribusi atau pemisahan kelas yang jelas antara data yang layak dan tidak layak minum.

3.3 Data *Preprocessing*

Pada *preprocessing* ini akan dilakukan seleksi data. Seleksi data yaitu melakukan pengecekan setiap variabel apakah ada data yang tidak lengkap atau tidak terisi. Data yang sudah *valid* kelengkapannya akan terpilih dan memasuki proses pengolahan data. *Preprocessing* yaitu proses pengolahan data yang terdapat nilai yang kosong (*missing value)*, data yang tidak lengkap ataupun adanya ketidaksesuaian data. Sehingga berdampak pada performa model dan hasil akurasinya.

Berdasarkan kondisi data pada *Exploratory Data Analysis* (EDA), maka dapat diselesaikan melalui *preprocessing* yaitu sebagai berikut:

3.3.1 Penghapusan Baris Data yang Kosong

Data kualitas air ini akan dilakukan *preprocessing* untuk diseleksi dan dihapus data-data yang terdapat nilai yang kosong serta data yang duplikat. Data-

data yang kosong tersebut akan diseleksi dan dihapus sehingga menjadi data yang siap untuk diolah dan diproses dengan algoritma KNN. Berikut hasil dari seleksi data pada data kualitas air setelah proses penghapusan baris data ditunjukkan pada Tabel 3.4.

Tabel 3.4 Data kualitas air setelah penghapusan baris data

No.	Ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic Carbon	Trihalome-thanes	Turbidity	Potability
1.	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
2.	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
3.	5.584087	188.313324	28748.687739	7.544869	326.678363	280.467916	8.399735	54.917862	2.559708	0
4.	10.223862	248.071735	28749.716544	7.513408	393.663396	283.651634	13.789695	84.603556	2.672989	0
5.	8.635849	203.361523	13672.091764	4.563009	303.309771	474.607645	12.363817	62.798309	4.401425	0
...	...	...	...	...	...	...	...	...	...	...
2007.	8.989900	215.047358	15921.412018	6.297312	312.931022	390.410231	9.899115	55.069304	4.613843	1
2008.	6.702547	207.321086	17246.920347	7.708117	304.510230	329.266002	16.217303	28.878601	3.442983	1
2009.	11.491011	94.812545	37188.826022	9.263166	258.930600	439.893618	16.172755	41.558501	4.369264	1
2010.	6.069616	186.659040	26138.780191	7.747547	345.700257	415.886955	12.067620	60.419921	3.669712	1
2011.	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1

Berdasarkan Tabel 3.4 pada *preprocessing* ini menghasilkan data yang *clear* berjumlah 2.011 data, dimana dari kondisi data mentah yang berjumlah 3.276 data menjadi 2.011 data yang siap diolah dan dapat dianalisis pada proses KNN selanjutnya.

3.3.2 Normalisasi *Min Max*

Normalisasi max min adalah salah bentuk normalisasi dengan untuk mempertahankan bentuk dan nilai pasti dari data minimum dan maksimum (Bu'ulolo, 2024). Implementasi proses normalisasi pada penelitian ini menggunakan teknik *Min Max Scaling*, yang mana membagi selisih antara nilai variabel terbesar dan nilai variabel terkecil dengan seluruh nilai variabel yang telah dikurangi nilai variabel terkecil. Tujuan dari dilakukannya normalisasi ini untuk mendapatkan nilai rentang, sehingga mampu menghasilkan data yang memiliki nilai piksel dengan rentang antara 0 dan 1. Perhitungan untuk proses normalisasi menggunakan *Min Max Scaling* ditunjukkan pada persamaan 3.1 (Pramana et al., 2020).

$$y = \frac{x - Dmin}{Dmax - Dmin} \tag{3.1}$$

Keterangan :
y = Nilai yang dinormalisasi (nilai rentang 0-1)
x = Nilai asli dalam data
Dmax = Nilai terbesar dalam data
Dmin = Nilai terkecil dalam data

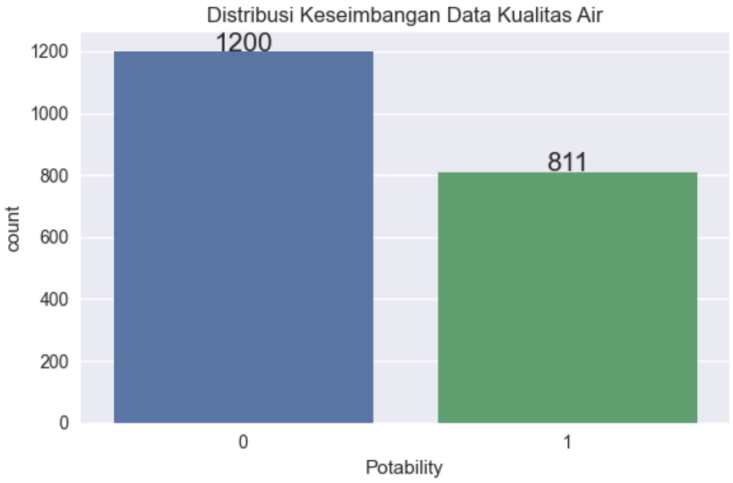
Berikut adalah hasil proses normalisasi yang telah diimplementasikan menggunakan Min Max Scaling yang dapat dilihat pada Tabel 3.5.

Tabel 3.5 Hasil Normalisasi *Min Max*

No.	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic Carbon	Trihalomethanes	Turbidity	Potability
1.	0.59	0.58	0.39	0.57	0.65	0.29	0.65	0.80	0.63	0
2.	0.64	0.44	0.31	0.44	0.51	0.36	0.38	0.20	0.52	0
3.	0.39	0.47	0.51	0.52	0.56	0.14	0.25	0.40	0.22	0
4.	0.73	0.72	0.51	0.52	0.75	0.15	0.47	0.66	0.24	0
5.	0.61	0.53	0.24	0.27	0.50	0.49	0.41	0.47	0.59	0
...	...	...	...	...	...	...	...	...	...	...
2006.	0.64	0.58	0.28	0.42	0.52	0.34	0.31	0.40	0.63	1
2007.	0.47	0.55	0.30	0.54	0.50	0.23	0.57	0.18	0.40	1
2008.	0.82	0.09	0.66	0.67	0.37	0.43	0.56	0.29	0.58	1
2009.	0.42	0.46	0.46	0.54	0.62	0.39	0.40	0.45	0.44	1
2011.	0.32	0.49	0.84	0.49	0.66	0.59	0.47	0.50	0.59	1

3.3.3 Penyeimbangan Data dengan SMOTE

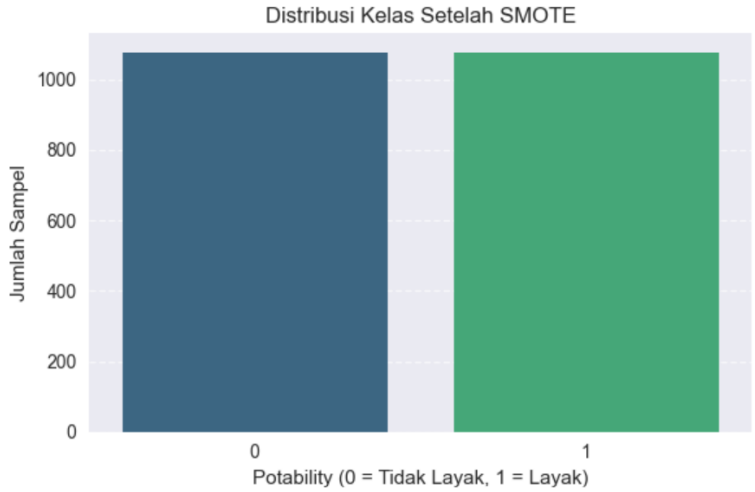
Setelah dilakukan penghapusan data yang kosong dan normalisasi data, Data yang digunakan dalam penelitian ini berjumlah 2.011 sampel, yang terdiri dari 1.200 data menunjukkan bahwa air tersebut tidak layak untuk diminum, sedangkan 811 sisanya merupakan data air yang masih tergolong aman dan layak dikonsumsi. Sehingga terjadi ketidakseimbangan kelas dengan proporsi air tidak layak minum lebih tinggi dibandingkan air yang layak. Distribusi ketidakseimbangan kelas pada Data Kualitas Air dapat dilihat pada gambar berikut.



Gambar 3.23 Distribusi Keseimbangan Data Kualitas Air

Grafik pada gambar 3.23 menampilkan jumlah sampel air berdasarkan kategorisasi potabilitas yang menunjukkan ketidakseimbangan kelas. Jumlah sampel air tidak layak minum (*potability* = 0) sebanyak 1.200 sampel data. Jumlah sampel air layak minum (*potability* = 1) sebanyak 800 sampel data. Maka dilakukan *preprocessing* dengan *SMOTE (Synthetic Minority Over-sampling Technique)*.

SMOTE (*Synthetic Minority Over-sampling Technique*) merupakan teknik *oversampling* cerdas yang bekerja dengan membuat data sintetis baru dari kelas minoritas (bukan hanya duplikasi). Teknik ini menggunakan teknik interpolasi yaitu membentuk data baru di antara titik-titik minoritas yang sudah ada. Teknik juga dapat mengurangi risiko *overfitting* akibat duplikasi. Hasil proses menyeimbangkan kelas dengan SMOTE dapat dilihat pada gambar 3.24.



Gambar 3.24 Distribusi Kelas Setelah SMOTE

Setelah penerapan SMOTE, gambar 3.24 menunjukkan jumlah data pada kedua kelas menjadi seimbang yaitu jumlah sampel air tidak layak minum (*potability* = 0) sebanyak 1.077 sampel data dan air layak minum (*potability* = 1) sebanyak 1.077 sampel data, sehingga total seluruh data kualitas air adalah 2.154 data. Ini memungkinkan model klasifikasi meningkatkan kemampuannya dalam mendeteksi kelas minoritas dan mengurangi bias terhadap kelas mayoritas.

3.4 *Data Training* dan *Data Testing*

Pada tahap ini, data akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih digunakan untuk membentuk atau melatih model agar bisa mengenali pola, sedangkan data uji dipakai untuk mengecek dan menilai seberapa baik kinerja model tersebut saat diterapkan pada data yang belum pernah dilihat sebelumnya. Pembagian data akan dilakukan pada skenario uji coba dengan beberapa pembagian rasio data.

3.5 *K-Nearest Neighbor*

Pada proses KNN ini akan dilakukan dengan menemukan model klasifikasi untuk mendapatkan akurasi tertinggi dengan memlakukan pengujian data *training* dan data *testing*. Pada sample yang digunakan dalam penelitian ini berupa data kualitas air yang terdiri dari 2.154 sample data dan 10 variabel.

Data kualitas air telah melalui tahap *preprocessing* dengan menghapus data-data yang kosong dan data yang duplikat, kemudian dilakukan normalisasi *Min Max* untuk mendapatkan nilai rentang antara 0 dan 1, serta telah dilakukan *SMOTE* yang dapat menyeimbangkan kelas pada data kualitas air minum sehingga proporsi kelas air tidak layak minum dan air yang layak minum seimbang. Selanjutnya data diuji dengan algoritma KNN dengan perhitungan jarak Manhattan dan akan diurutkan berdasarkan nilai *k* dengan tujuan untuk mendapatkan model. Sehingga akan didapatkan model klasifikasi yang dijabarkan dalam bentuk tabel *Confusion Matrix* untuk menampilkan hasil pengujian. Kemudian hasilnya akan disesuaikan dengan hasil dari data aslinya.

Pada sample yang digunakan dalam pengujian data menggunakan algoritma K-Nearest Neighbor diambil secara acak, berupa data kualitas air dengan 15 sample data dan 10 variabel. Berikut sampel data latih serta data uji ditunjukkan pada Tabel 3.6 Sampel Data Training Kualitas Air.

Tabel 3.6 Sampel Data Training Kualitas Air

Data Training										
No.	**ph**	**Hardness**	**Solids**	**Chloramines**	**Sulfate**	**Conductivity**	**Organic Carbon**	**Trihalomethanes**	**Turbidity**	**Potability**
1.	0.42	0.63	0.18	0.39	0.51	0.43	0.61	0.49	0.68	0
2.	0.29	0.19	0.61	0.56	0.80	0.43	0.49	0.49	0.56	0
3.	0.55	0.45	0.24	0.46	0.59	0.23	0.53	0.72	0.65	0
4.	0.44	0.46	0.73	0.70	0.67	0.57	0.38	0.58	0.58	0
5.	0.65	0.82	0.42	0.47	0.77	0.50	0.45	0.54	0.61	0
6.	0.41	0.46	0.78	0.25	0.81	0.24	0.53	0.44	0.80	1
7.	0.51	0.41	0.60	0.69	0.37	0.42	0.57	0.63	0.60	1
8.	0.55	0.34	0.41	0.66	0.56	0.17	0.60	0.55	0.41	1
9.	0.80	0.02	0.69	0.61	0.59	0.38	0.49	0.39	0.37	1
10.	0.65	0.19	0.47	0.40	0.61	0.32	0.35	0.40	0.68	1
11.	0.39	0.60	0.15	0.59	0.49	0.38	0.17	0.62	0.34	0
12.	0.61	0.43	0.32	0.50	0.65	0.23	0.53	0.65	0.32	0
13.	0.62	0.50	0.43	0.64	0.57	0.24	0.58	0.51	0.47	1
14.	0.50	0.50	0.30	0.47	0.54	0.58	0.43	0.50	0.51	1
15.	0.35	0.65	0.45	0.39	0.57	0.59	0.46	0.26	0.43	1
Data Uji										
1.	0.66	0.51	0.43	0.40	0.57	0.35	0.58	0.19	0.57	?

Data latih yang ditampilkan pada Tabel 3.6 akan digunakan sebagai acuan untuk mengelompokkan data uji berdasarkan jarak terdekatnya. Data uji yang dimaksud merupakan data baru yang sebelumnya belum termasuk dalam kelompok mana pun. Data ini diambil dari hasil *preprocessing*, lalu akan diuji dan diklasifikasikan menggunakan model yang sudah dibangun.

3.5.1 *Manhattan Distance*

Selanjutnya adalah mengklasifikasikannya ke dalam salah satu dari dua sifat air yang dapat diminum atau kelas *Potability* dengan mengukur perbedaan absolut antara nilai numerik dari setiap fitur dengan perhitungan *Manhattan distance* terhadap k mayoritas tetangga terdekat. Perhitungan *Manhattan distance* dilakukan berdasarkan rumus perhitungan jarak *Manhattan distance* pada persamaan (2.1). Perhitungan jarak menggunakan Manhattan Distance ditampilkan pada Tabel 3.7. Nilai-nilai tersebut diperoleh dari hasil menghitung jarak antara data uji dengan data latih, yang digunakan sebagai dasar dalam proses klasifikasi.

Tabel 3.7 Hasil Perhitungan *Manhattan distance*

No.	Manhattan Distance	Potability	Ranking
1.	1.21	0	5
2.	1.72	0	13
3.	1.25	0	6
4.	1.77	0	14
5.	1.26	0	7
6.	1.69	1	12
7.	1.44	1	10
8.	1.30	1	9
9.	1.65	1	11
10.	0.99	1	2
11.	2.01	0	15
12.	1.29	0	8
13.	0.84	1	1
14.	1.13	1	4
15.	1.05	1	3

3.5.2 Klasifikasi Data

Setelah jarak antara data uji dan data latih dihitung menggunakan rumus Manhattan, langkah selanjutnya adalah mengurutkan data latih yang paling dekat dengan data uji. Kemudian, data tersebut diklasifikasikan ke dalam kelas *Potability* berdasarkan mayoritas dari nilai k yang digunakan. Pengujian ini dilakukan dengan mencoba beberapa nilai k, yaitu k = 3, k = 5, dan k = 7, untuk melihat hasil klasifikasinya. Hasil data uji yang sesuai dengan ketetanggaan terdekat yaitu diberi nilai 1 pada kelas *Potability* yang berarti kualitas air "baik" atau dapat dikonsumsi dan nilai 0 yang artinya kualitas air "buruk" atau tidak dapat dikonsumsi. Berikut hasil uji pada urutan jarak *Manhattan* berdasarkan nilai k=3 ditampilkan di Tabel 3.8.

Tabel 3.8 Urutan jarak *Manhattan* berdasarkan nilai k=3

No.	k=3		
	Manhattan Distance	**Potability**	**Ranking**
1.	0.84	1	1
2.	0.99	1	2
3.	1.05	1	3

Berdasarkan hasil uji dengan k=3 pada Tabel 3.8, mayoritas jarak terpendek terhadap label k adalah *Potability* dengan nilai 1 yaitu kualitas air "baik". Maka data uji masuk ke dalam kelompok sifat air yang aman dikonsumsi berdasarkan *Manhattan Distance*. Berikut hasil uji pada urutan jarak *Manhattan* berdasarkan nilai k=5 ditampilkan di Tabel 3.9.

Tabel 3.9 Urutan jarak *Manhattan* berdasarkan nilai k=5

No.	k=5		
	Manhattan Distance	**Potability**	**Ranking**
1.	0.84	1	1
2.	0.99	1	2
3.	1.05	1	3
4.	1.13	1	4
5.	1.21	0	5

Berdasarkan hasil uji dengan k=5 pada Tabel 3.9, Hasil pengujian menunjukkan bahwa sebagian besar data yang memiliki jarak terdekat terhadap nilai k mengarah pada label *Potability* dengan nilai 1, yang berarti kualitas air tersebut dikategorikan sebagai "baik". Maka data uji termasuk ke dalam kelompok sifat air yang aman dikonsumsi berdasarkan *Manhattan Distance*. Berikut hasil uji pada urutan jarak *Manhattan* berdasarkan nilai k=7 ditampilkan di Tabel 3.10.

Tabel 3.10 Urutan jarak *Manhattan* berdasarkan nilai k=7

No.	k=7		
	Manhattan Distance	**Potability**	**Ranking**
1.	0.84	1	1
2.	0.99	1	2
3.	1.05	1	3
4.	1.13	1	4
5.	1.21	0	5
6.	1.25	0	6
7.	1.26	0	7

Berdasarkan hasil uji dengan k=7 pada Tabel 3.10, mayoritas jarak terpendek terhadap label k adalah *Potability* dengan nilai 1 yaitu kualitas air "baik". Maka data uji termasuk ke dalam kelompok sifat air yang aman dikonsumsi berdasarkan *Manhattan Distance*.

Dari hasil pengujian dengan nilai k = 3, k = 5, dan k = 7, menunjukkan mayoritas jarak terpendek terhadap kelas *Potability* yang bernilai 1, maka dapat

disimpulkan sampel baru (data uji) diklasifikasikan sebagai kualitas air "Baik" yaitu sifat air aman untuk dikonsumsi.

3.6 Evaluasi Model

Langkah akhir dalam penelitian ini adalah melakukan evaluasi terhadap performa model klasifikasi guna mengukur tingkat akurasi algoritma KNN. Evaluasi dilakukan dengan menggunakan *confusion matrix* sebagai alat analisis, yang memungkinkan perhitungan sejumlah metrik evaluasi, yaitu akurasi, presisi (*precision*), sensitivitas (*recall*), dan skor F1 (*F1-Score*). Rumus perhitungan untuk masing-masing metrik tersebut telah disajikan pada Tabel 2.4 sebagai acuan dalam proses evaluasi model.

3.7 Skenario Uji Coba

Dalam mengevaluasi kinerja model yang telah dibuat, peneliti melakukan beberapa skenario pengujian menggunakan perbandingan data dan perbandingan nilai *k*. Skenario pengujian ini dilakukan untuk melihat seberapa besar pengaruhnya terhadap tingkat akurasi dari model. Setelah hasil prediksi didapatkan, langkah selanjutnya adalah menghitung nilai akurasi, presisi, *recall*, dan *f1-score* dengan menggunakan metode *confusion matrix* sebagai alat evaluasinya.

3.7.1 Skenario Perbandingan Data

Pada skenario pengujian pertama, data dibagi menjadi dua bagian, yaitu data latih dan data uji, dengan menggunakan beberapa rasio perbandingan yang berbeda. Data latih digunakan untuk membangun model, sedangkan data uji berfungsi untuk mengecek seberapa baik model tersebut bekerja. Mengacu pada penelitian

sebelumnya oleh (Kartini et al., 2022) pendekatan serupa digunakan dalam memprediksi penyakit menggunakan algoritma KNN, dengan rasio pembagian data 90:10, 80:20, dan 70:30, untuk melihat bagaimana pengaruhnya terhadap hasil akurasi, presisi, recall, dan F1-score. Rincian skenario pembagian data dalam penelitian ini bisa dilihat pada Tabel 3.11.

Tabel 3.11 Skenario perbandingan data

No.	Data *Training*	Data *Testing*
1.	90%	10%
2.	80%	20%
3.	70%	30%

Tabel 3.11 merupakan tabel perbandingan data *training* dan data *testing* untuk dilakukan uji coba sistem. Penelitian dilakukan dengan membagi rasio data untuk mengetahui rasio mana yang memiliki tingkat akurasi tertinggi.

3.7.2 Skenario Perbandingan Nilai *k*

Pada skenario uji coba yang kedua yaitu melakukan perbandingan nilai *k*. Nilai *k* digunakan untuk mengurutkan data latih berdasarkan kedekatan jarak *Manhattan* dengan data uji, guna menentukan mayoritas kelas dari tetangga terdekat yang menjadi dasar klasifikasi. Mengacu pada studi terdahulu oleh (Puteri et al., 2023) membandingkan kinerja KNN dengan nilai k = 3, 5, dan 7 dalam mendiagnosis penyakit dimana variasi nilai *k* pada algoritma KNN mempengaruhi akurasi dalam kinerja klasifikasi. Skenario perbandingan nilai *k* ditunjukkan di Tabel 3.12.

Tabel 3.12 Skenario perbandingan nilai k

No.	Nilai Parameter *k*
1.	k=3
2.	k=5
3.	k=7

BAB IV

UJI COBA DAN PEMBAHASAN

Pada bagian ini akan dijelaskan hasil pengujian sistem berdasarkan beberapa skenario uji coba yang telah dilakukan, dengan menggunakan *confusion matrix* sebagai dasar evaluasi untuk memperoleh nilai akurasi, presisi (*precision*), sensitivitas (*recall*), dan skor F1 (*F1-score*). Dalam mengevaluasi kinerja model yang telah dibuat, peneliti melakukan beberapa skenario pengujian menggunakan perbandingan data dan perbandingan nilai *k*. Seluruh pembahasan dalam bab ini mengacu pada skenario pengujian yang sudah dijelaskan sebelumnya di sub-bab 3.8.

4.1 Hasil Uji Coba

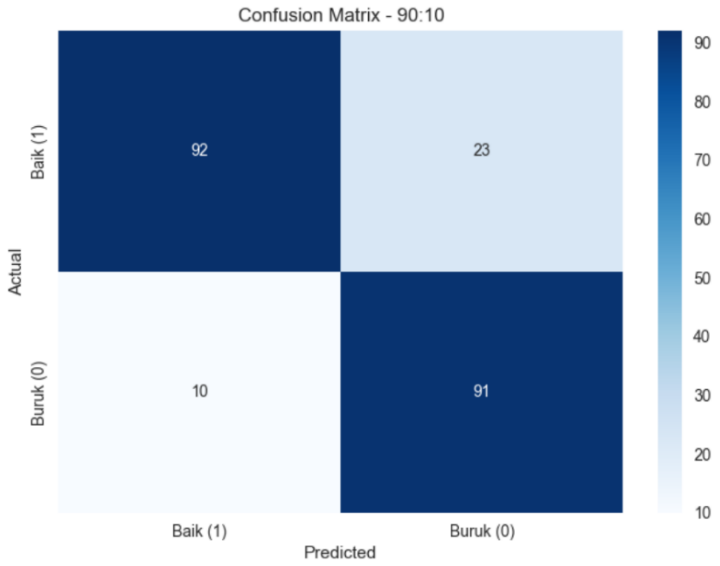
4.1.1 Uji Coba Rasio Data 90:10

Skenario uji coba yang pertama menggunakan rasio data 90:10 dengan nilai k tetangga terdekat 3, 5, dan 7 dimana pengujian algoritma KNN ini menggunakan perbandingan 90% *data training* sebanyak 1.938 data kualitas air dan 10% *data testing* sebanyak 216 data kualitas air. Hasil *confusion matrix* yang didapatkan pada rasio 90:10 untuk *recall*, akurasi, presisi, dan *f1-score* yang ditampilkan di Tabel 4.1.

Tabel 4.1 Nilai *Accuracy*, *Precision*, *Recall*, *F1-Score* pada Rasio Data 90:10

Pengujian Nilai K	Rasio Data 90:10			
	Accuracy	*Precision*	*Recall*	*F1-Score*
3	85%	90%	80%	85%
5	81%	85%	79%	82%
7	72%	78%	67%	72%
Rata-rata	79%	84%	75%	80%

Dari Tabel 4.1 terlihat bahwa nilai akurasi paling tinggi tercapai saat menggunakan k = 3, yaitu 85%. Ini menunjukkan bahwa sistem sangat mampu dalam memprediksi kualitas air secara akurat. Hasil klasifikasi dengan k = 3 ini kemudian digambarkan lebih jelas melalui *confusion matrix* yang ditampilkan pada Gambar 4.1.



Gambar 4.1 *Confusion Matrix* 90:10 dengan nilai k=3

Dalam dataset angka “1” sebagai kualitas air baik dan angka “0” sebagai kualitas buruk. Berdasarkan hasil *confusion matrix*, diketahui bahwa ada 92 data yang berhasil diprediksi dengan benar oleh sistem sebagai kualitas air “baik” atau layak konsumsi. Artinya, sistem mampu mengenali dan mengklasifikasikan data tersebut secara tepat ke dalam kategori yang seharusnya yaitu *True Positive*, 10 data FP (*False Positive*) yaitu data “Buruk” yang salah diklasifikasikan sebagai data “Baik”, 23 data FN (*False Negative*) yang salah memprediksi kualitas air “buruk” tetapi kenyataan diprediksi “baik”, dan 91 data TN (*True Negative*) sistem berhasil memprediksi data dengan benar sebagai kualitas air “buruk”.

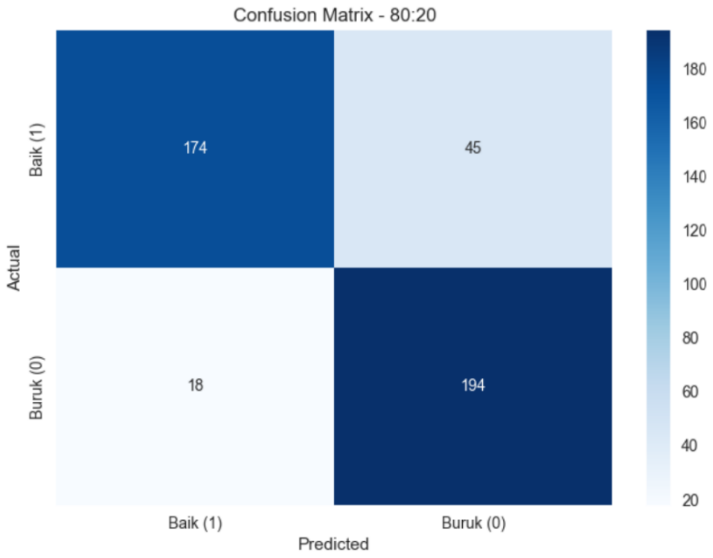
4.1.2 Uji Coba Rasio Data 80:20

Skenario pengujian data selanjutnya menggunakan rasio data 80:20 dengan nilai k tetangga terdekat 3, 5, dan 7 dimana pengujian algoritma KNN ini menggunakan perbandingan 80% *data training* sebanyak 1.723 data kualitas air dan 20% *data testing* sebanyak 431 data kualitas air. Hasil *confusion matrix* yang didapatkan pada rasio 80:20 untuk *recall*, akurasi, presisi, dan *f1-score* yang ditampilkan di Tabel 4.2.

Tabel 4.2 Nilai *Accuracy, Precision, Recall, F1-Score* pada Rasio Data 80:20

Pengujian Nilai K	Rasio Data 80:20			
	Accuracy	*Precision*	*Recall*	*F1-Score*
3	85%	91%	79%	85%
5	82%	85%	79%	82%
7	75%	78%	72%	75%
Rata-rata	81%	85%	77%	81%

Dari Tabel 4.2 terlihat bahwa nilai akurasi paling tinggi tercapai saat menggunakan k = 3, yaitu 85%. Ini menunjukkan bahwa sistem cukup mampu dalam memprediksi kualitas air secara akurat. Hasil klasifikasi dengan k = 3 ini kemudian digambarkan lebih jelas melalui *confusion matrix* yang ditampilkan pada Gambar 4.2.



Gambar 4. 2 *Confusion Matrix* 80:20 pada k=3

Berdasarkan hasil evaluasi menggunakan *confusion matrix*, diperoleh data sebagai berikut: sebanyak 174 sampel tergolong sebagai *True Positive* (TP), dimana sistem berhasil secara akurat mengklasifikasikan kualitas air "baik". 18 data FP (*False Positive*) yaitu Data kualitas air "buruk" yang salah diprediksi sebagai data "baik". Selanjutnya, terdapat 45 sampel yang tergolong *False Negative* (FN), yakni kualitas air "baik" yang salah diprediksi sebagai "buruk". Terakhir, sebanyak 194 sampel diklasifikasikan secara tepat sebagai kualitas air "buruk", yang masuk ke dalam kategori *True Negative* (TN).

4.1.3 Uji Coba Rasio Data 70:30

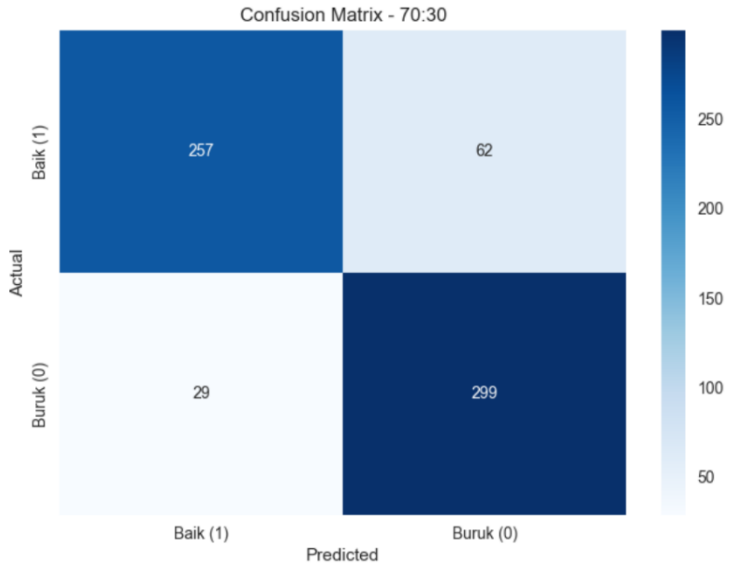
Skenario uji coba yang ketiga menggunakan rasio data 70:30 dengan nilai k tetangga terdekat 3, 5, dan 7 dimana pengujian algoritma KNN ini menggunakan perbandingan 70% *data training* sebanyak 1.507 data kualitas air dan 30% *data testing* sebanyak 647 data kualitas air. Hasil evaluasi model berdasarkan *confusion matrix* pada rasio data 70% untuk pelatihan dan 30% untuk pengujian dapat dilihat

pada Tabel 4.3. Tabel tersebut menampilkan nilai akurasi, presisi, *recall*, dan *f1-score* yang diperoleh dari pengujian dengan rasio tersebut.

Tabel 4.3 Nilai *Accuracy, Precision, Recall, F1-Score* pada Rasio Data 70:30

Pengujian Nilai K	Rasio Data 70:30			
	Accuracy	*Precision*	*Recall*	*F1-Score*
3	86%	90%	81%	85%
5	81%	83%	78%	81%
7	75%	76%	71%	74%
Rata-rata	81%	83%	77%	80%

Pada Tabel 4.3 menunjukkan akurasi yang paling tinggi diperoleh saat menggunakan jumlah k=3 yaitu akurasi 86% yang artinya sistem dapat memprediksi kualitas air dengan baik. Selanjutnya nilai k=3 pada *confusion matrix* ditampilkan di Gambar 4.3 sebagai berikut.



Gambar 4.3 *Confusion Matrix* 70:30 pada k=3

Berdasarkan hasil dari confusion matrix, sistem berhasil memprediksi dengan benar sebanyak 257 data kualitas air yang tergolong "baik" (*True Positive*)., 29 data FP (*False Positive*) yaitu Data kualitas air "buruk" yang salah diprediksi

sebagai data “baik”, Sementara itu, terdapat 62 data yang sebenarnya masuk kategori “baik”, namun salah diprediksi oleh sistem sebagai “buruk” (*False Negative*), dan 299 data TN (*True Negative*) sistem berhasil memprediksi data dengan benar sebagai kualitas air “buruk”.

4.2 Pembahasan

Uji coba telah dilakukan menggunakan data sebanyak 2.154 data kualitas air minum dengan 10 variabel. Sebelumnya, data ini terlebih dahulu diolah pada proses seleksi data pada tahap *preprocessing* yaitu menghapus data-data yang terdapat nilai kosong, kemudian proses normalisasi untuk menghasilkan data dengan nilai rentang antara 0 dan 1 menggunakan *Min Max Scaling*, lalu terakhir menyeimbangkan kelas data dengan *SMOTE*. Sehingga menghasilkan sebanyak 2.154 data kualitas air minum untuk siap diproses sistem. Kemudian hasil data tersebut digunakan pada beberapa uji coba KNN menggunakan variasi rasio data dan nilai k yang berbeda.

Hasil pada data kualitas air minum tersebut akan diproses dengan klasifikasi algoritma K-Nearest Neighbor (KNN), menggunakan variasi rasio data dan nilai k yang berbeda. Pembagian rasio antara data latih dan data uji yang digunakan dalam pengujian tersebut dapat dilihat pada Tabel 4.4.

Tabel 4.4 Perbandingan data latih terhadap data uji

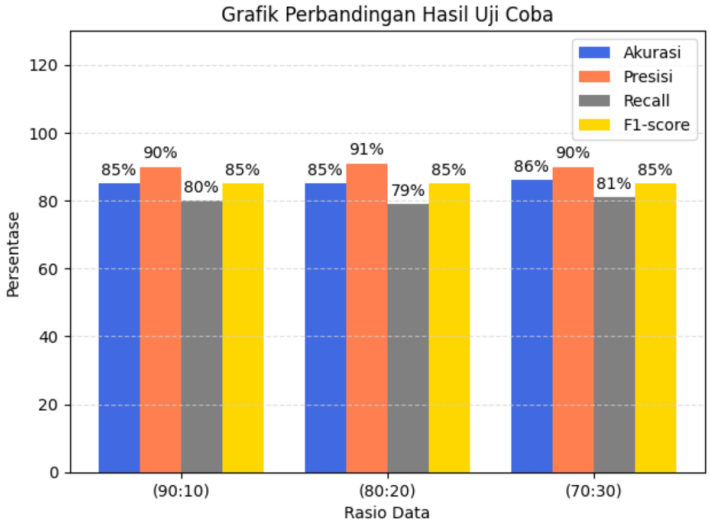
Data Latih : Data uji (%)	Data latih	Data uji
90 : 10	1.938	216
80 : 20	1.723	431
70 : 30	1.507	647

Pada pengujian data menggunakan algoritma KNN, maka akan didapatkan jarak *Manhattan* yang berdasarkan jarak ketetanggaan terdekat akan diurutkan untuk mendapatkan model. Hasil pengujian data uji terhadap data *training* terhadap 15 sampel data dengan nilai k=3, 5, dan 7 telah dibahas pada sub-bab 3.5.2 Klasifikasi Data. Berdasarkan hasil pengujian tersebut, menunjukkan bahwa mayoritas jarak terpendek terhadap label k adalah kelas *Potability* dengan nilai 1 sehingga diklasifikasikan sebagai kategori kualitas air "baik". Data uji tersebut termasuk ke dalam kelompok sifat air yang aman dikonsumsi berdasarkan *Manhattan Distance*. Maka dapat disimpulkan, model yang dihasilkan pada pengujian *K-Nearest Neighbor* dapat memprediksikan kelas *Potability* yang bernilai 1 dengan lebih baik dibandingkan kelas *Potability* yang bernilai 0.

Hasil pengujian sistem dengan beberapa skenario uji coba menggunakan perbandingan data yaitu rasio 80:20, 90:10, dan 70:30 serta perbandingan jumlah k=3, 5, dan 7 akan ditunjukkan pada grafik perbandingan hasil uji coba.

4.2.1 Grafik Perbandingan Hasil Uji Coba berdasarkan Nilai k=3

Berikut grafik perbandingan hasil uji coba pada rasio 90:10, 8020, dan 70:30 berdasarkan nilai k=3 dapat dilihat pada gambar 4.4.

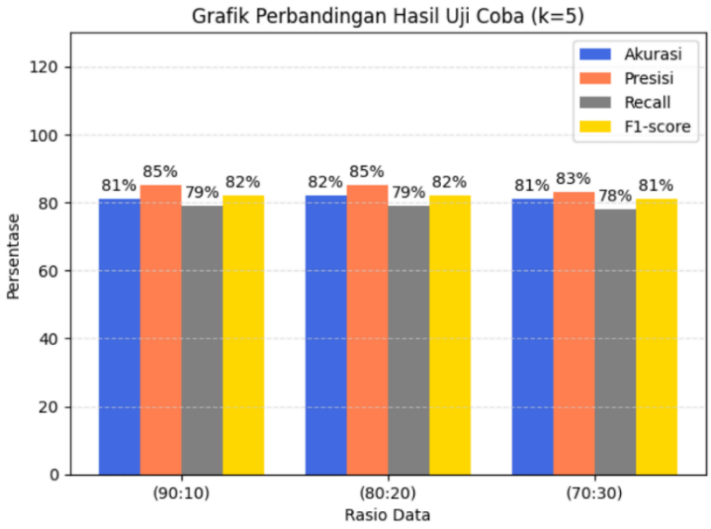


Gambar 4.4 Grafik Perbandinga Hasil Uji Coba berdasarkan nilai k=3

Terlihat dari Gambar 4.4 berdasarkan nilai k=3 bahwa akurasi tertinggi diperoleh oleh rasio 70:30 dimana akurasi 86%, presisi 90%, *recall* sebessar 81% serta *F1-Score* 85%. Sedangkan pada rasio 80:20 memiliki presisi yang lebih tinggi yang menunjukkan model sangat berhati-hati ketika memutuskan nilai positif (minim kesalahan positif palsu). Rasio 90:10 memiliki pola yang sama dengan rasio lainnya dimana nilai presisi yang paling tinggi yaitu 90%.

4.2.2 Grafik Perbandingan Hasil Uji Coba berdasarkan Nilai k=5

Berikut grafik perbandingan hasil uji coba pada rasio 90:10, 8020, dan 70:30 berdasarkan nilai k=5 dapat dilihat pada gambar 4.5.

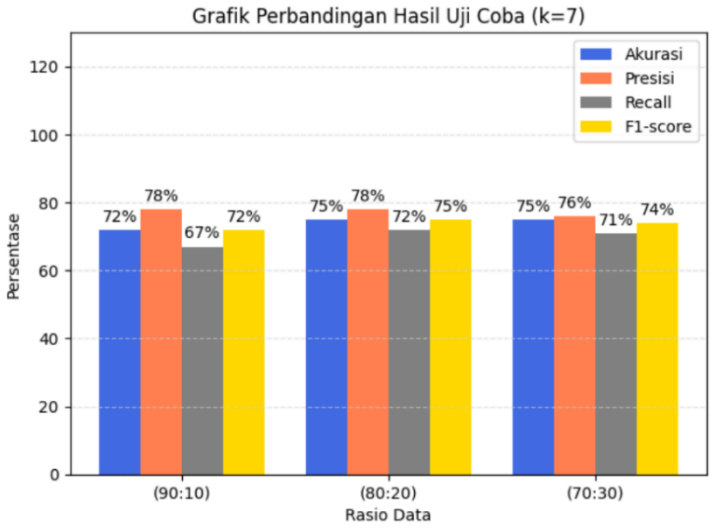


Gambar 4.5 Grafik Perbandinga Hasil Uji Coba berdasarkan nilai k=5

Terlihat dari Gambar 4.5 berdasarkan nilai k=5 bahwa akurasi tertinggi diperoleh oleh rasio 80:20 dimana akurasi 82%, presisi 85%, *recall* sebessar 79% serta *F1-Score* 82%. Dari seluruh hasil grafik rasio data, model cenderung konsisten namun performanya lebih rendah daripada grafik sebelumnya dengan nilai k=3. Hasil grafik k=5 memperlihatkan bahwa presisi memiliki nilai lebih tinggi dibandingkan *recall* yang menunjukkan bahwa model cenderung menghindari prediksi positif palsu, nemun terkadang melewatkan data positif.

4.2.3 Grafik Perbandingan Hasil Uji Coba berdasarkan Nilai k=7

Berikut grafik perbandingan hasil uji coba pada rasio 90:10, 8020, dan 70:30 berdasarkan nilai k=7 dapat dilihat pada gambar 4.6.



Gambar 4 6 Grafik Perbandinga Hasil Uji Coba berdasarkan nilai k=7

Terlihat dari Gambar 4.6 berdasarkan nilai k=7 bahwa akurasi tertinggi diperoleh oleh rasio 80:20 dan rasio 70:30 dengan nilai akurasi sebesar 75%, dimana presisi, *recall,* dan *F1-Score* dari kedua rasio data ini memiliki nilai yang hamper sama. Namun grafik pada nilai k=7 ini cenderung lebih rendah daripada grafik pada nilai k=5 dan k=3, sehingga pada nilai k=7 model kurang optimal karena tidak seimbang antara *recall* dan presisi serta performanya lebih rendah.

Ringkasan hasil terbaik dari pemodelan yang telah dilakukan bisa dilihat pada tabel 4.5 berikut. Nilai-nilai tersebut diperoleh dari perbandingan antara jumlah data latih dan data uji, serta variasi jumlah k yang digunakan selama proses pengujian.

Tabel 4.5 Performa Pengujian Nilai Tertinggi

Rasio Data (%)	**Jumlah nilai k (ketetanggaan)**	***Accuracy***	***Precision***	***Recall***	***F1-Score***
90 : 10	3	85%	90%	80%	85%
80 : 20	3	85%	91%	79%	85%
70 : 30	3	86%	90%	81%	85%

Berdasarkan tabel 4.5 Dari hasil penelitian mengenai klasifikasi kualitas air minum, kombinasi antara data latih dan data uji dengan perbandingan 70%:30% serta penggunaan nilai k sebesar 3 memberikan hasil terbaik. Pada pengujian ini, model berhasil mencapai akurasi sebesar 86%, dengan presisi 90%, *recall* 81%, dan *f1-score* 85%.

Nilai akurasi tersebut menunjukkan bahwa sistem mampu mengklasifikasikan data dengan benar sebanyak 86% dari seluruh data uji yang digunakan. Presisi 90% menunjukkan bahwa 90% data benar-benar sesuai dengan seluruh data yang diprediksi sebagai kualitas “baik”, sehingga model sangat baik dalam meminimalisir kesalahan (*false positive*). *Recall* 81% menunjukkan bahwa Dari seluruh data yang sebenarnya berkualitas "Baik", hanya 81% yang berhasil dikenali dengan benar. Sedangkan 85% pada *f1-score* memberikan gambaran keseimbangan antara kemampuan model dalam mengidentifikasi kualitas air yang “baik” secara akurat dan menyeluruh. Berdasarkan hasil pengujian tertinggi pada performa model pada rasio data 70:30 dengan k=3 termasuk dalam kategori sangat baik.

Jumlah ketetanggaan nilai k pada pengujian tertinggi adalah nilai k=3. Pengujian tertinggi bisa terjadi pada nilai k yang terkecil disebabkan karena ketika nilai k=3 model bisa lebih fokus ke tetangga terdekat yang benar-benar sama dengan data uji, sehingga hasil prediksi kualitas air lebih tepat sasaran. Berbeda dengan k=5 dan k=7 dimana model terlalu banyak melihat jumlah tetangga yang berbeda-beda dan belum tentu sama dengan data uji, sehingga hasil prediksi yang dihasilkan kurang akurat.

Berdasarkan hasil pengujian, bisa disimpulkan bahwa sistem mampu bekerja dengan baik dalam mengelompokkan kualitas air minum ke dalam dua kategori, yaitu "baik" (layak diminum) dan "buruk" (tidak layak diminum), dengan bantuan algoritma *K-Nearest Neighbors*. Dalam penelitian ini, tingkat akurasi yang dihasilkan sangat dipengaruhi oleh beberapa hal, seperti seberapa besar rasio antara data latih dan data uji, serta berapa jumlah nilai k yang digunakan dalam proses klasifikasi.

Dari serangkaian uji coba yang telah dilakukan, digunakan beberapa pembagian data latih dan data uji dengan rasio 90:10, 80:20, dan 70:30 serta skenario uji coba dengan perbandingan nilai k={3, 5, 7}, pengujian tertinggi diperoleh dengan akurasi tertinggi yaitu rasio data 70:30 dengan nilai k=3. Maka dapat disimpulkan, semakin banyak *data training* dan semakin kecil jumlah nilai *k* yang digunakan dalam penggunaan algoritma *K-Nearest Neighbor* dalam pengelompokan kualitas air minum terbukti mampu memberikan tingkat akurasi yang semakin baik.

Berdasarkan hasil penelitian, sistem ini dapat digunakan untuk membantu proses klasifikasi, khususnya dalam memperkirakan apakah air layak dikonsumsi atau tidak. Hal ini juga sejalan dengan firman Allah Swt. dalam ayat 21 pada Surah Al-Hijr yaitu:

وَاِنْ مِّنْ شَيْءٍ اِلَّا عِنْدَنَا خَزَاۤىِٕنُهٗ وَمَا نُنَزِّلُهٗٓ اِلَّا بِقَدَرٍ مَّعْلُوْمٍ ﴿٢١﴾

"Tidak ada sesuatu pun melainkan di sisi Kami-lah perbendaharaannya dan Kami tidak menurunkannya melainkan dengan ukuran tertentu." (QS. Al-Hijr ayat 21)

Menurut tafsir Jalalain, ayat ini menjelaskan bahwa semua hal yang ada di dunia ini sebenarnya berasal dari perbendaharaan milik Allah, dan semuanya diturunkan dengan takaran yang pas sesuai dengan kebutuhan. Kalimat "min" dalam ayat tersebut hanya sebagai penekanan tambahan, sementara maksud utamanya adalah bahwa segala sesuatu ada kuncinya di sisi Allah. Ini menunjukkan bahwa setiap ciptaan Allah tidak muncul begitu saja, melainkan sudah diatur dengan sangat teliti dan penuh kebijaksanaan (As-suyuti, 2025). Maka, dengan adanya sistem klasifikasi ini diharapkan mampu melakukan klasifikasi sesuai ukurannya yang akurat, seperti Allah menciptakan sesuatu sesuai dengan ukurannya. Salah satu tujuan dilakukannya pada data penelitian yaitu untuk mendapatkan ukuran yang baik dengan cara pengujian dengan rasio pembagian data dan pemilihan nilai k untuk memperoleh Tingkat akurasi yang baik dalam kinerja klasifikasi. Dengan cara ini, kesalahan dalam pengambilan keputusan bisa diminimalkan karena informasi yang digunakan untuk mengelompokkan data jadi lebih akurat. Hal ini sangat penting dalam proses pengembangan aplikasi, supaya hasil yang dihasilkan benar-benar sesuai dengan kondisi sebenarnya dan tidak menyesatkan. Salah satunya sistem dapat memprediksi kualitas air dan dapat membantu memudahkan masyarakat maupun tenaga perusahaan air minum mendapatkan informasi.

Saat ini, sebagian besar sumber air untuk kebutuhan air minum di Indonesia berasal dari air permukaan seperti sungai, danau, dan waduk. Sayangnya, banyak dari sumber air tersebut sudah tidak sejernih dulu karena tercemar oleh berbagai

aktivitas manusia. Contohnya seperti pembuangan sampah sembarangan, limbah rumah tangga, dan kegiatan industri yang tidak ramah lingkungan, yang akhirnya merusak kualitas air bersih. Pencemaran air dapat berdampak pada kehidupan manusia yang berpengaruh pada kesehatannya. Dalam ajaran Islam, hal ini juga ditegaskan oleh Allah Swt. melalui salah satu ayat-Nya. Seperti yang tercantum dalam Al-Qur'an, tepatnya pada ayat 205 Surah Al-Baqarah yaitu:

وَإِذَا تَوَلَّىٰ سَعَىٰ فِي الْأَرْضِ لِيُفْسِدَ فِيهَا وَيُهْلِكَ الْحَرْثَ وَالنَّسْلَ ۗ وَاللَّهُ لَا يُحِبُّ الْفَسَادَ

"Dan apabila ia berpaling (dari kamu), ia berjalan di bumi untuk mengadakan kerusakan padanya, dan merusak tanam-tanaman dan binatang ternak, dan Allah tidak menyukai kebinasaan." (QS. Al-Baqarah ayat 205).

Menurut penjelasan tafsir Jalalayn, ayat tersebut menjelaskan bahwa ada orang-orang yang ketika berpaling dari ajaran yang benar, mereka justru berjalan di muka bumi untuk menyebarkan kerusakan. Kerusakan yang dimaksud bisa berupa penghancuran tanaman dan pembunuhan hewan ternak. Tindakan seperti ini menunjukkan bentuk-bentuk kerusakan yang nyata, padahal Allah tidak menyukai segala bentuk kerusakan di muka bumi. Artinya, Allah tidak meridhai perbuatan tersebut karena bertentangan dengan nilai-nilai yang diajarkan dalam Islam (As-suyuti, 2025). Segala bentuk kerusakan lingkungan, termasuk mencemari air, tanah, dan udara, adalah bentuk fasad (kerusakan) yang dilarang dalam Islam. Islam secara prinsip memerintahkan manusia untuk menjaga sumber air dan melarang pencemarannya, baik karena dapat menyebabkan kerusakan (fasad), bertentangan dengan *maqāṣid syarī'ah*, dan berdampak pada kesehatan, ibadah, dan kehidupan makhluk lainnya. Maka dari itu kita dianjurkan untuk menjaga dan melestarikan

sumber air yang berarti menjaga kehidupan, sedangkan mencemari air adalah bentuk pelanggaran terhadap amanah Allah sebagai khalifah di bumi. Berdasarkan tujuan dari penelitian ini yaitu untuk mendeteksi air yang berkualitas, maka sistem ini dibuat agar membantu menjalankan perintah Allah untuk menjaga sumber air yang dapat dikonsumsi oleh masyarakat.

Terkait dengan menjalankan perintah Allah dalam memudahkan masyarakat untuk menjaga kualitas air minum, dengan membuat sistem yang dapat memprediksi kualitas air minum, yaitu hukumnya *Sunnah muakkadah* (sangat dianjurkan) karena memudahkan akses masyarakat terhadap air yang bersih dan berkualitas. Sebagaimana Hadits Riwayat Abu Dawud dan Ahmad yang berbunyi : Rasulullah ﷺ bersabda:

اَلْمُسْلِمُوْنَ شُرَكَاءُ فِي ثَلَاَثٍ فِي الْكَلَإِ وَالْمَاءِ وَالنَّار

"Manusia itu berserikat dalam tiga perkara: air, padang rumput, dan api." (HR Abu Dawud dan Ahmad)

Dalam kitab *al-Mabsûth*, Imam as-Sarakhsyi menyebutkan bahwa air merupakan milik bersama yang haknya tidak boleh dimiliki pribadi. Tidak ada satu pun orang yang berhak melarang orang lain untuk menggunakan air, karena air tidak boleh dimonopoli atau dibuat sulit untuk diakses (PRATAMA, 2019). Maka memudahkan orang lain mendapat air bersih adalah perbuatan baik yang sangat dianjurkan. Dalam Islam, memudahkan akses masyarakat terhadap air yang bersih dan berkualitas adalah tindakan yang sangat dianjurkan, bahkan wajib dalam

kondisi darurat atau saat tidak ada alternatif lain. Ini termasuk dalam nilai kemanusiaan, pelayanan sosial, dan *amar ma'ruf*.

Sebagai seorang muslim, kita diperintahkan oleh Allah untuk mengonsumsi makanan dan minuman yang tidak hanya halal tapi juga baik bagi tubuh. Setiap muslim hukumnya wajib (*fardhu*) mencari makanan dan minuman yang halal serta berkualitas, karena hal itu termasuk kewajiban dalam agama. Kewajiban ini menjadi bagian dari menjaga agama, jiwa, dan kesehatan menurut *maqāṣid al-syarī'ah* (tujuan syariat) dan ditegaskan pada Al-Qur'an pada ayat 172 Surah Al-Baqarah:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا كُلُوْا مِنْ طَيِّبٰتِ مَا رَزَقْنٰكُمْ وَاشْكُرُوْا لِلّٰهِ اِنْ كُنْتُمْ اِيَّاهُ تَعْبُدُوْنَ

"Wahai orang-orang yang beriman, makanlah dari rezeki yang baik yang kami berikan kepada kamu dan bersyukurlah kepada Allah Swt, jika kamu hanya menyembah-Nya." (QS. Al-Baqarah : 172).

Berdasarkan penelitian yang dilakukan oleh (Samsudin, 2020) bahwa dalam Tafsir Al-Azhar menurut Hamka dijelaskan bahwa manusia dianjurkan untuk mengonsumsi makanan yang halal dan baik bagi tubuh (thoyyib). Anjuran ini lebih ditujukan kepada orang-orang beriman karena makanan bisa memengaruhi kondisi perilaku dan jiwa seseorang. Oleh karena itu, kita dianjurkan untuk berusaha mendapatkan air yang berkualitas dengan tujuan air tersebut baik saat diminum.

BAB V

PENUTUP

5.1 Kesimpulan

Setelah mencoba berbagai pembagian antara data latih dan data uji (90:10, 80:20, dan 70:30) serta menggunakan beberapa nilai k (3, 5, dan 7), hasil terbaik didapat saat menggunakan 70% data latih dan k = 3. Pada kondisi ini, akurasinya mencapai 86%, presisinya 90%, *recall*-nya 81%, dan *F1-score*-nya 85%. Jadi bisa disimpulkan bahwa perbandingan jumlah data latih dan data uji serta pemilihan nilai k memberi pengaruh besar terhadap seberapa baik model ini bekerja dan seberapa akurat hasilnya. Semakin banyak data yang dipakai untuk melatih model dan semakin kecil nilai k yang dipakai, maka hasil akurasinya menjadi lebih baik. Algoritma KNN (*K-Nearest Neighbors*) mampu memprediksi kualitas air minum dengan baik sehingga menghasilkan *output* nilai parameter klasifikasi dengan tingkat akurasi yang cukup baik dalam mengidentifikasi kualitas air minum yang layak dikonsumsi atau tidak.

5.2 Saran

Saran yang dapat dilakukan pada penelitian selanjutnya yaitu:

1. Menggunakan *K-Fold Cross Validation* pada proses evaluasi model sehingga performa model dapat lebih akurat karena metode ini dapat memilih model terbaik dan dapat menghindari *overfitting*.

2. Terlihat dari hasil *recall* pada pengujian kualitas air minum dengan KNN, maka dapat dilakukan perbandingan performa dengan uji model lain seperti *Random forest* atau *Logistic Regression.*

DAFTAR PUSTAKA

Addzikri, A. I., & Rosariawari, F. (2023). Analisis Kualitas Air Permukaan Sungai Brantas Berdasarkan Parameter Fisik dan Kimia. *INSOLOGI: Jurnal Sains Dan Teknologi*, *2*(3), 550–560. https://doi.org/10.55123/insologi.v2i3.1981

As-suyuti, D. A. N. J. (2025). *KAJIAN KITAB TAFSIR AL-JALALAIN KARYA JALALUDDIN AL MAHALLI*. *2*, 211–227.

Bu'ulolo, E. (2024). ALGORITMA K-NEAREST NEIGHBOR (K-NN) DENGAN NORMALISASI MAX MIN UNTUK MENENTUKAN CALON MAHASISWA YANG LAYAK MENERIMA KIP KULIAH MERDEKA. *Jurnal Sistem Informasi Dan Sistem Komputer*, *9*(2), 184–194.

Harahap, P. N., & Sulindawaty, S. (2020). Implementasi Data Mining Dalam Memprediksi Transaksi Penjualan Menggunakan Algoritma Apriori (Studi Kasus PT.Arma Anugerah Abadi Cabang Sei Rampah). *Matics*, *11*(2), 46. https://doi.org/10.18860/mat.v11i2.7821

Hartanti, D., & Pradana, A. I. (2023). Komparasi Algoritma Machine Learning dalam Identifikasi Kualitas Air. *SMARTICS Journal*, *9*(1), 1–6. https://doi.org/10.21067/smartics.v9i1.8113

Jaelani, A. (2020). *Deteksi Dan Klasifikasi Tipe Bangunan Pada Citra Satelir Menggunakan Metode K-Nearest Neighbor*.

Kartini, D., Farmadi, A., Muliadi, M., Turianto Nugrahadi, D., & Pirjatullah, P. (2022). Perbandingan Nilai K pada Klasifikasi Pneumonia Anak Balita Menggunakan K-Nearest Neighbor. *Jurnal Komputasi*, *10*(1), 47–53. https://doi.org/10.23960/komputasi.v10i1.2965

Kualitas, K., & Menggunakan, A. I. R. (n.d.). *Klasifikasi kualitas air menggunakan metode extreme learning machine (elm) 1,2, 3*. 983–994.

Lobo, A. C. (2022). Tinjauan Yuridis Terhadap Dampak Pencemaran Air Terhadap Kesehatan Masyarakat Di Desa Poponcol Kabupaten Karawang. *JUSTITIA: Jurnal Ilmu Hukum Dan Humaniora*, *9*(3), 1386–1394. http://jurnal.um-tapsel.ac.id/index.php/

Malik, F., & Akbar, N. (2024). Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air. *Jurnal TEKNO KOMPAK*, *18*(1), 28–40. https://ejurnal.teknokrat.ac.id/index.php/teknokompak/article/view/3241/1412

Mardiana, L., Kusnandar, D., & Satyahadewi, N. (2022). Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)*, *11*(1), 97–102.

Muhammad, C., Maulana, R., & Ichsan, M. H. H. (2020). Purwarupa Perahu untuk Monitoring dan Klasifikasi Kualitas Air Bendungan dengan Metode K-Nearest Neighbor (KNN). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(2), 651–659. http://j-ptiik.ub.ac.id

Musdalifah. (2022). ANALISIS KUALITAS AIR DAN BEBAN PENCEMARAN BERDASARKAN PARAMETER FISIKA DAN KIMIA DI DANAU UNIVERSITAS HASANUDDIN. *Repository Universitas Hasanuddin*. http://repository.unhas.ac.id:443/id/eprint/18276

Mutoffar, M. M., & Fadillah, A. (2022). Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest. *Naratif : Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, *4*(2), 138–146. https://doi.org/10.53580/naratif.v4i2.160

Na'imah, S. (2021). *Berbagai Masalah Kesehatan Akibat Pencemaran Air di Indonesia*. HelloSehat. https://hellosehat.com/sehat/informasi-kesehatan/pencemaran-air-sebab-dan-dampak-kesehatan/

Nada Osseiran & Yemi Lufadeju. (2019). *1 in 3 people globally do not have access to safe drinking water – UNICEF, WHO*. World Health Organitation. https://www.who.int/news/item/18-06-2019-1-in-3-people-globally-do-not-have-access-to-safe-drinking-water-unicef-who

Nurmahaludin, & Cahyono, G. R. (2019). Klasifikasi Kualitas Air PDAM Menggunakan Algoritma KNN Dan K-Means. *Prosiding SNRT*, *5662*(November), 1–7.

Pangestu, M. S., & Fitriani, M. A. (2022). Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means. *Sainteks*, *19*(2), 141. https://doi.org/10.30595/sainteks.v19i2.14495

Pramana, I. G. B. B. A., Widiartha, I. M., & Astuti, L. G. (2020). Implementation Learning Vector Quantization(LVQ) for Chronic Kidney Disease Classification. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, *9*(2), 241. https://doi.org/10.24843/jlk.2020.v09.i02.p11

PRATAMA, I. R. (2019). *TINJAUAN HUKUM ISLAM TENTANG PRAKTIK JUAL BELI AIR DARI SUMBER AIR MILIK BERSAMA (Studi Kasus di Kecamatan Gisting Kabupaten Tanggamus)*. 1–57.

Prihambodo, B., Wildan F Y, A., Prayoga, E., & Jaffar, A. (2023). Klasifikasi Kualitas Air Sungai Berbasis Teknik Data Mining Dengan Metode K-Nearest Neighbor (K-NN). *Emitor: Jurnal Teknik Elektro*, *23*(1). https://doi.org/10.23917/emitor.v1i1.20833

Puteri, Q. A., Sagirani, T., & Lemantara, J. (2023). Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) untuk Mengetahui Keakuratan Diagnosa Penyakit Diabetes. *Jurnal Nasional Teknologi Dan Sistem*

Informasi, *9*(3), 247–254. https://doi.org/10.25077/teknosi.v9i3.2023.247-254

Rachmat, R., Chrisnanto, Y., & Fajri Rakhmat Umbara. (2020). Sistem Prediksi Mutu Air Di Perusahaan Daerah Air Minum Tirta Raharja Menggunakan K–Nearest Neighbors (K–NN). *Seminar.Iaii.or.Id*, *2*(7), 915–922. http://www.seminar.iaii.or.id/index.php/SISFOTEK/article/view/211

Said, H., Matondang, N. H., & Irmanda, H. N. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi. *Techno.Com*, *21*(2), 256–267. https://doi.org/10.33633/tc.v21i2.5901

Samsudin. (2020). Makanan Halal Dan Thayyib Perspektif Al-Qur'an. *Book Chapter*, 1–26.

Saputra, H. M., Sari, M., Purnomo, T., Suhartawan, B., Asnawi, I., Palupi, I. F. J., Sahabuddin, E. S., Sinaga, J., Juhanto, A., Yuniarti, E., & Nur, S. (2023). *ANALISIS KUALITAS LINGKUNGAN* (Issue August). https://www.researchgate.net/publication/374118371

Sarker, I. H., Faruque, F., Alqahtani, H., & Kalim, A. (2020). K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(26), 1–9. https://doi.org/10.4108/eai.13-7-2018.162737

Sopiatul Ulum, Alifa, R. F., Rizkika, P., & Rozikin, C. (2023). Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum. *Generation Journal*, *7*(2), 141–146. https://doi.org/10.29407/gj.v7i2.20270

Sutisna, & Yuniar, N. M. (2023). Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes. *Jurnal Sains Dan Teknologi*, *5*(1), 243–246. https://doi.org/10.55338/saintek.v5i1.1383

Swantika, I. M. A., Kanata, B., & Suksmadana, I. M. B. (2020). Perancangan Sistem Untuk Mengetahui Kualitas Biji Kopi Berdasarkan Warna Dengan K-Nearest Neighbor. *Jurnal Bakti Nusa*, *1*(2), 25–36.

Tangkelayuk, A., & Evangs, M. (2022). Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(2), 1109–1119. https://doi.org/10.35957/jatisi.v9i2.2048

Vidiastanta, I. G., Hidayat, N., & Dewi, R. K. (2020). Komparasi Metode K-Nearest Neighbors (K-NN) Dengan Support Vector Machine (SVM) Untuk Klasifikasi Status Kualitas Air. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(1), 312–319. http://j-ptiik.ub.ac.id

Zamri, N., Pairan, M. A., Azman, W. N. A. W., Abas, S. S., Abdullah, L., Naim, S., Tarmudi, Z., & Gao, M. (2022). River quality classification using different distances in k-nearest neighbors algorithm. *Procedia Computer Science*, *204*(2021), 180–186. https://doi.org/10.1016/j.procs.2022.08.022