

Klasifikasi Kualitas dan Prediksi Kondisi Air Tanah di DKI Jakarta Menggunakan Algoritma Naïve Bayes

Arsya Fathiarahma¹, Nina Sulistiyowati¹, Taufik Ridwan¹, dan Apriade Voutama¹

¹Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang, Karawang, Indonesia

Corresponding author: Arsyafathiarahma (e-mail: arsyafathiarahma19005@student.unsika.ac.id)

ABSTRACT Groundwater remains an important source for meeting the daily clean water needs of most DKI Jakarta residents, with 32% of the population relying on this source. High population density and associated environmental issues, such as the lack of adequate water catchment and sanitation infrastructure, have triggered concerns over groundwater quality in the region. In particular, the contamination of groundwater by bacteria such as *E. coli* presents serious risks to public health such as stomach cramps, the presence of blood during diarrhea, and vomiting. This research aims to address these issues by analyzing groundwater quality in different areas of DKI Jakarta. This approach involves visual mapping using QGIS to illustrate groundwater pollution, thus facilitating a deeper understanding of groundwater conditions. The total data of 1068 used in this study was obtained from two different sources and analyzed using the Naïve Bayes Gaussian classification algorithm. The evaluation technique involved the use of cross-validation and percentage split in 10 scenarios. The results showed that the highest accuracy rate of 84.36% as well as precision values of 0.8566, recall 0.8436, and F1-score 0.8436 were obtained using the percentage split technique with a ratio of 80:20. The findings can increase public awareness of groundwater quality in DKI Jakarta and contribute to more effective planning of water resources protection efforts.

KEYWORDS Classification, Effective, Groundwater, Naïve Bayes Gaussian

ABSTRAK Air tanah tetap menjadi sumber penting untuk pemenuhan kebutuhan air bersih sehari-hari bagi sebagian besar masyarakat DKI Jakarta, dengan 32% penduduknya mengandalkan sumber ini. Kepadatan penduduk yang tinggi dan masalah lingkungan yang terkait, seperti kurangnya infrastruktur resapan air dan sanitasi yang memadai, telah memicu keprihatinan akan kualitas air tanah di wilayah ini. Terutama, terkontaminasinya air tanah oleh bakteri seperti *E. coli* menghadirkan risiko serius bagi kesehatan masyarakat seperti perut kram, adanya darah saat diare, hingga muntah-muntah. Penelitian ini bertujuan untuk mengatasi masalah tersebut dengan menganalisis kualitas air tanah di berbagai wilayah DKI Jakarta. Pendekatan ini melibatkan pemetaan visual menggunakan QGIS untuk mengilustrasikan pencemaran air tanah, sehingga memfasilitasi pemahaman yang lebih mendalam mengenai kondisi air tanah. Total data sejumlah 1068 yang digunakan dalam penelitian ini diperoleh dari dua sumber yang berbeda dan dianalisis menggunakan algoritma klasifikasi *Naïve Bayes Gaussian*. Teknik evaluasi melibatkan penggunaan *cross-validation* dan *percentage split* dalam 10 skenario. Hasil penelitian menunjukkan tingkat akurasi tertinggi sebesar 84.36% serta nilai *precision* sejumlah 0.8566, *recall* 0.8436, dan *F1-score* 0.8436 diperoleh menggunakan teknik *percentage split* dengan rasio 80:20. Temuan ini dapat meningkatkan kesadaran masyarakat tentang kualitas air tanah di DKI Jakarta dan berkontribusi pada perencanaan upaya perlindungan sumber daya air yang lebih efektif.

KATA KUNCI Air tanah, Efektif, Klasifikasi, *Naïve Bayes Gaussian*

I. PENDAHULUAN

Air adalah suatu senyawa kimia yang sangat dibutuhkan untuk keberlangsungan hidup di bumi. Sebagian besar wilayah permukaan bumi, sebanyak 72% ditutupi oleh air. Pernyataan tersebut dikuatkan oleh penelitian yang dilakukan oleh agensi ilmiah pemerintah Amerika Serikat, USGS (United States Geological Survey). Menurut USGS, jumlah volume air di bumi sekitar 332.500.000 mil kubik atau 1.4 juta km³ dengan sumber air yang terdiri dari air laut, air permukaan, air atmosfer, air sungai, dan air tanah[1]

Air disebut sebagai sumber kehidupan dilandasi dengan banyaknya kebutuhan yang dapat terpenuhi diantaranya adalah proses fotosintesis yang dilakukan oleh tumbuhan serta sumber minum dan tempat tinggal untuk hewan. Selain itu, bergantungnya kepada air dirasakan oleh manusia untuk keperluan sehari-hari seperti mandi, memasak, mencuci dan lain-lain[2]. Standar kelayakan kebutuhan air bersih berjumlah 49,5 liter per kapita per hari dan untuk kebutuhan tubuh manusia diperlukan sekiranya 2.5 liter/hari, jumlah tersebut tidaklah kecil, sehingga perlu adanya pemilihan sumber air yang memadai. Jenis air tanah dangkal atau air sumur adalah salah satu sumber air untuk memenuhi kebutuhan tersebut[3].

DKI Jakarta merupakan salah satu daerah yang sebanyak 32% masyarakat masih memanfaatkan air tanah untuk memenuhi kebutuhan sehari-hari. Menurut ahli Hidrologi dari Universitas Brawijaya, Dr Gunawan Wibisono, perlu adanya sistem pengelolaan air yang baik dan teratur serta pengawasan dan penelusuran sumber air tanah yang akan dikonsumsi. Buruknya sanitasi di DKI Jakarta, membuat air limbah seperti septic tank mudah terkontaminasi dengan bakteri, salah satunya bakteri *e.coli*. Bakteri *Escherichia coli* (*E. coli*) adalah salah satu jenis bakteri yang menyebabkan infeksi usus pada manusia dengan menghasilkan racun yang disebut verotoksin sebagai penyebab kerusakan dinding usus kecil dan menyebabkan gejala seperti perut kram, adanya darah saat diare, hingga muntah-muntah[4]

Data mining merupakan sebuah proses mengekstrak informasi dan wawasan yang berguna dari data yang besar dan kompleks. Proses ini menggunakan teknik-teknik analisis data yang canggih untuk mengidentifikasi pola, hubungan, dan tendensi yang mungkin tidak terlihat secara langsung dari data tersebut. Guna menganalisa atau memperkirakan data hasil ukur selanjutnya mengenai kualitas air tanah di DKI Jakarta, dapat dilakukan proses penambangan data dari data sebelumnya. Tahapan pengolahan data melibatkan beberapa langkah, seperti pengumpulan data awal, pembersihan data dari *outlier* atau data yang tidak valid, transformasi data ke format yang sesuai, dan pemilihan atribut atau fitur yang relevan untuk analisis.

Selain dilakukannya penambahan data, untuk meningkatkan kesadaran masyarakat terhadap kualitas air tanah di DKI Jakarta, penelitian ini bertujuan memberikan informasi yang dapat disebarluaskan kepada masyarakat

dengan visualisasi data kualitas air tanah yang diintegrasikan dalam bentuk peta interaktif, sehingga informasi tersebut dapat lebih mudah dipahami dan diakses oleh masyarakat umum. Dengan adanya pemaparan mengenai kualitas air tanah serta kandungan kimia atau bakteri yang berakibat fatal bagi tubuh, diharapkan masyarakat mencegah dampak negatif tersebut pada kesehatan, menjadi lebih sadar akan pentingnya menjaga kualitas air tanah dan dapat mengambil tindakan yang sesuai untuk melindungi sumber daya air yang begitu berharga ini.

Penelitian terdahulu yang dilakukan oleh M. Ja'far Sodik pada tahun 2019 mengatakan bahwa klasifikasi kualitas udara di DKI Jakarta yang dilakukan dengan metode *Naive Bayes* mendapatkan nilai *accuracy* sebesar 91.862 % [5] Selain itu, penelitian lain dilakukan oleh Dani Purwanto dalam kasus klasifikasi *naive Bayes* untuk memprediksi kualitas udara berdasarkan Indeks Standar Pencemar Udara (ISPU)[6] yang memiliki *accuracy* sebesar 93%. Kemudian, penelitian yang dilakukan oleh Aditya Rizky Sanjaya di tahun 2019 dengan judul klasifikasi jenis mobil paling diminati di Indonesia dengan menggunakan algoritma *naive Bayes* mendapatkan *accuracy* 96,7[7]. Namun, perlu diperhatikan bahwa penelitian sebelumnya belum adanya fokus pada kualitas air tanah dan penggunaan algoritma *Naive Bayes Gaussian* untuk tujuan tersebut. Oleh karena itu, tidak dapat dilakukan perbandingan rinci dengan penelitian sebelumnya. Kombinasi antara kualitas air tanah dan algoritma *Naive Bayes Gaussian* ini merupakan hal yang baru dan inovatif, yang diharapkan dapat memberikan wawasan berharga mengenai kualitas air tanah di DKI Jakarta.

Berdasarkan dari beberapa penelitian sebelumnya, metode *naive Bayes* merupakan metode yang memiliki *accuracy* cukup tinggi. Kelebihan penggunaan algoritma *Naive Bayes Gaussian* adalah kemampuannya untuk mengatasi data yang memiliki sifat kontinu [8], seperti konsentrasi bahan kimia dalam air tanah. Algoritma ini juga dikenal karena sederhananya, yang memungkinkan pemrosesan cepat dan efisien. Selain itu, *Naive Bayes* biasanya memerlukan jumlah data pelatihan yang lebih kecil dibandingkan dengan beberapa algoritma lainnya, yang sangat bermanfaat karena data kualitas air tanah di DKI Jakarta terbatas dalam jumlahnya [9]. Dengan dilakukannya beberapa skema pada tahap pemodelan, diharapkan dapat menghasilkan model dengan akurasi terbaik dalam klasifikasi kualitas air tanah di DKI Jakarta

II. TINJAUAN PUSTAKA

A. DATA MINING

Data mining merupakan suatu proses yang memanfaatkan teknik dan algoritma komputasi untuk mengolah data secara masif dalam jumlah yang sangat besar dan bervariasi[10]. *Data mining* bermula dari pertemuan antara berbagai disiplin ilmu seperti *machine learning*, *pattern recognition*, *artificial intelligence*, dan sistem basis data, yang semuanya memiliki

tujuan untuk menangani data yang sangat besar, heterogen, dan tersebar [11]

Data mining sangat berguna dalam menyelesaikan masalah yang dihadapi oleh perusahaan, organisasi, atau individu yang mengumpulkan data secara masif dan ingin mengetahui pola atau tren dari data tersebut[12]. Dengan menggunakan data mining, kita dapat menemukan pola-pola yang tidak terlihat secara visual dari data yang ada, sehingga dapat memberikan informasi yang berguna bagi pengambilan keputusan atau perencanaan strategis[13]. *Data mining* juga dapat digunakan untuk mengoptimalkan sistem informasi, meningkatkan efisiensi proses bisnis, dan meningkatkan keuntungan[14]

B. NAÏVE BAYES

Naïve bayes merupakan salah satu metode populer dalam pengklasifikasian probabilitas sederhana untuk keperluan data mining karena dengan struktur yang cukup sederhana, metode ini dapat mudah diimplemantisasikan serta tingginya efektifitas yang dihasilkan[15]. Metode yang dikemukakan oleh ilmuwan Inggris Thomas Bayes ini dikenal sebagai teorema bayes yaitu memprediksi peluang kedepannya, dengan menggunakan pengalaman sebelumnya sebagai bahan perhitungan[16]. Arti naif dari metode ini adalah karena membuat asumsi kemunculan fitur tidak saling bergantung dan arti tersebut merupakan ciri dari *naïve bayes*, yaitu independent[17]. Proses klasifikasi pada metode *naïve bayes* memerlukan sebuah data set yang terdiri dari *data training* atau dapat disebut data latih dan *data testing* atau dapat disebut data uji yang kemudian akan diproses dan membentuk pengetahuan berupa nilai probabilitas pada setiap data[18]. Metode *naïve bayes* bertujuan untuk melakukan klasifikasi data pada kelas tertentu dengan kelebihan dapat menangani data baik yang bersifat diskrit maupun kontinu[8]. Pada atribut data yang bersifat kontinu, metode yang digunakan yaitu *naïve bayes gaussian*, dimana Persamaan (1) sebagai metode *naïve bayes gaussian*.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

C. CONFUSION MATRIX MULTICLASS

Confusion matrix multiclass merupakan perkembangan dari *matriks confusion binary*, di mana sebelumnya terdapat FN (*False Negative*), FP (*False Positive*), dan TN (*True Negative*). Namun, pada *matriks confusion multiclass*, hanya terdapat TP karena untuk menentukan FN adalah dari jumlah baris keseluruhan per variabel, sedangkan untuk menentukan FP adalah dari jumlah kolom keseluruhan per variabel, dan TN adalah kasus-kasus di mana prediksi kita tidak ada dan nilai aktualnya salah.

Confusion matriks multiclass adalah tabel yang umumnya digunakan untuk menjelaskan kinerja model klasifikasi pada dataset pengujian tertentu, di mana nilai sebenarnya sudah diketahui[19]. Contoh *confusion matriks* dapat dilihat pada Gambar 1

		Predicted				
		Label 1	Label 2	...	Label n	
Actual	Label 1	TP	E	...	E	Total Baris 1
	Label 2	E	TP	...	E	Total Baris 2

	Label n	E	E	...	TP	Total Baris n
		Total Kolom 1	Total Kolom 2	...	Total Kolom n	

GAMBAR 1. *Confusion matrix multiclass*

Hasil evaluasi *confusion matrix multiclass* berupa nilai *accuracy*, *precision* dan *recall*. *Accuracy* adalah jumlah proporsi prediksi benar dengan membandingkan dengan total seluruh data. Adapun rumus perhitungan yang digunakan untuk menghitung *accuracy* dapat dilakukan dengan Persamaan (2).

$$\frac{TP(\text{label } 1) + \dots + TP(\text{label } n)}{\text{total}(\text{baris } 1) + \dots + \text{total}(\text{baris } n)} \quad (2)$$

Recall adalah jumlah proporsi prediksi benar dengan membandingkan dengan total data positif. Adapun rumus perhitungan yang digunakan untuk menghitung *recall* dapat dilakukan dengan Persamaan (3).

$$\frac{TP(\text{label } i)}{\text{Total}(\text{baris } i)} \quad (3)$$

Precision adalah jumlah proporsi prediksi benar dengan membandingkan dengan yang diprediksi positif. Adapun rumus perhitungan yang digunakan untuk menghitung *precision* dapat dilakukan dengan Persamaan (4).

$$\frac{TP(\text{label } i)}{\text{Total}(\text{kolom } i)} \quad (4)$$

F1-score adalah rata-rata harmoni *precision* dan *recall* dengan menggambarkan keseluruhan kinerja model yang memberikan keseimbangan antara *precision* dan *recall*. Adapun rumus perhitungan yang digunakan untuk menghitung *F1-score* dapat dilakukan dengan Persamaan (5).

$$2 \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (5)$$

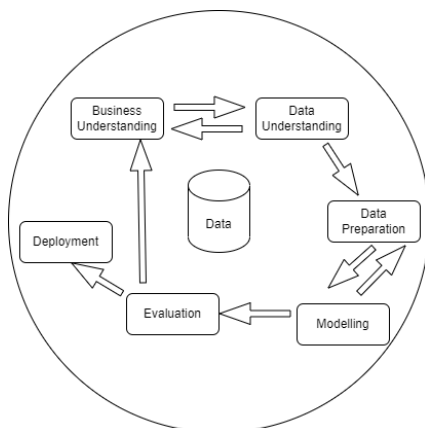
D. PERCENTAGE SPLIT

Percentage split adalah salah satu opsi pengujian yang tersedia dalam data mining. Hasil klasifikasi *percentage split* diuji dengan menggunakan bagian dari dataset sebagai *data training*. Dalam *percentage split*, data dibagi secara acak menjadi dua kelompok terpisah. Kelompok pertama digunakan untuk mengambil pengetahuan dari *data training*, dan pengetahuan yang diambil kemudian diuji pada kelompok kedua yang disebut sebagai data uji. Pembagian data acak dapat dilakukan dengan beberapa rasio yang disesuaikan dengan kebutuhan penelitian[20]

E. CROSS VALIDATION

K-Fold Cross validation merupakan metode statistik yang digunakan untuk evaluasi kinerja model klasifikasi dengan membagi data menjadi 2 bagian yaitu data uji dengan guna untuk proses pembelajaran untuk membentuk model klasifikasi dan data latih dengan guna melakukan validasi terhadap model tersebut. *K-Fold Cross validation*, data latih dibagi menjadi K bagian yang disebut "*fold*". Setiap *fold* kemudian digunakan secara bergantian sebagai data uji dan sisanya sebagai data latih. Proses ini dilakukan K kali, dengan setiap *fold* digunakan sekali sebagai data uji. Setelah *K-Fold Cross validation* selesai dilakukan, kita dapat menghitung rata-rata dari K hasil evaluasi yang diperoleh untuk mendapatkan estimasi kinerja model secara keseluruhan. Dengan menggunakan *K-Fold Cross validation*, kita dapat memperoleh estimasi yang lebih akurat dari kinerja model klasifikasi kita karena data latih yang lebih banyak digunakan untuk validasi[21]

III. METODOLOGI PENELITIAN



GAMBAR 2. Metodologi CRISP-DM

Metodologi yang dilakukan pada penelitian ini dipaparkan pada Gambar 2, yaitu metode CRISP-DM dengan enam proses, *bussines understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation* dan *deployment*.

A. BUSINESS UNDERSTANDING

Tahap ini akan dilakukannya pemahaman penelitian secara menyeluruh pada laporan akhir dari pemantauan kualitas air di DKI Jakarta sehingga mendapatkan tujuan yang akan dicapai pada penelitian ini. Kegiatan tersebut dilakukan terhadap data, serta dokumen yang telah disediakan Dinas Lingkungan Hidup DKI Jakarta dan dapat diunduh secara bebas oleh publik. Setelah dilakukannya pemahaman terhadap data dan dokumen yang ada, maka selanjutnya menyiapkan rencana dan strategi untuk mencapai tujuan tersebut.

B. DATA UNDERSTANDING

Tahap ini akan dilakukannya pengumpulan data yang didapat dari Dinas Lingkungan Hidup DKI Jakarta dan Open Data DKI Jakarta. Data tersebut dieksplor secara menyeluruh, kemudian dipelajari agar dapat menemukan wawasan awal atau informasi secara spesifik yang mempunyai karakteristik, sehingga akan didapatkannya kualitas dan kelengkapan data

C. DATA PREPARATION

Tahap ini akan dilakukannya suatu tahap intensif karena perlu adanya analisa data secara rinci agar hasil yang diinginkan dapat dicapai. Pengolahan data atau persiapan data dilakukan sebelum proses klasifikasi dilakukan dapat disebut preprocessing. Beberapa preprocessing yang akan dilakukan diantaranya pemilihan data dan atribut yang akan digunakan pada penelitian untuk proses klasifikasi, selanjutnya penggabungan data dari berbagai sumber yang kemudian dipadukan satu sama lain, melakukan format data agar data dapat diolah oleh tools yang digunakan, pemilihan feature dan label, dan preprocessing terakhir adalah normalisasi data

D. MODELLING

Tahap ini akan dilakukannya training dan testing terhadap data yang sudah dikumpul dan diolah. Digunakannya algoritma *naïve bayes* dengan dua teknik yang berbeda, *cross validation* dan *percentage split* serta bantuan tools google colab yang di dalamnya terdapat beberapa library python menjadikan adanya perbedaan hasil sebagai bahan perbandingan agar dapat memilih skenario dengan hasil terbaik

E. EVALUATION

Tahap ini dilakukannya interpretasi dari hasil pemodelan data mining. Berdasarkan tujuan yang telah dijelaskan pada tahap bussines understanding, maka pada tahap ini akan dilakukan Analisa hasil klasifikasi data prediksi kualitas air tanah. Evaluasi dilakukan dengan pengimplementasian *confusion matrix* dengan 2 teknik data mining, *cross validation* serta *percentage split* sehingga akan mendapatkan hasil yang telah direncanakan.

F. DEPLOYMENT

Tahap ini akan dilakukannya penulisan laporan hasil kegiatan yang sudah dilakukan dan tinjauan mengenai kualitas air tanah di DKI Jakarta pada tahun 2019 dengan 2 periode dan 2021 dengan 2 periode divisualisasikan dengan menggunakan bantuan tools QGIS. Pada laporan akhir didapatkan mengenai hasil analisis dari awal sampai hasil evaluasi yang mencakup *accuracy*, *precision*, *recall*, dan *f1-score*.

IV. HASIL DAN PEMBAHASAN

A. BUSINESS UNDERSTANDING

Pemahaman bisnis berdasarkan penelitian kualitas air tanah di DKI Jakarta pada tahun 2019 dan 2021 menghasilkan 3 poin utama diantaranya *determine business objectives, asses the situation, dan determine data mining goal.*

1) DETERMINE BUSINESS OBEJECTIVES

Tujuan bisnis dari penelitian ini berdasarkan data dan laporan akhir kualitas air tanah di DKI Jakarta yaitu mendapatkan hasil mengenai mutu air tanah di DKI Jakarta pada tahun 2019 dan 2021. Selain itu, peneletian ini mempunyai tujuan lain diantaranya dapat mengevaluasi hasil *accuracy* dari kualitas kondisi air tanah serta performa terkait *confusion matrix*. Parameter suksesnya prediksi kualitas kondisi air tanah dilihat dari beberapa atribut diantaranya kelurahan, kecamatan, nilai pH, deterjen, mangan, total coliform, e.coli, nilai indeks pencemaran dan status mutu.

2) ASESS THE SITUATION

Kondisi data awal yang diperoleh mempunyai bentuk yang berbeda karena didapatkan dari 2 sumber. Pada tahun 2019, sumber data diambil dari open data DKI Jakarta dan untuk tahun 2021, data diambil dari Dinas Lingkungan Hidup DKI Jakarta pada dokumen yang berjudul "Laporan Akhir Pemantauan Kualitas Air Tanah Tahun 2021". Data awal 2019 mempunyai 20 paramter sebagai bahan perhitungan indeks pencemaran dalam bentuk csv. Namun, pada tahun 2021, data diperoleh dari grafik dengan 5 parameter di dalamnya, diantaranya, pH, mangan, deterjen, total coliform, dan e.coli. Selain grafik, terdapat data berbentuk tabel yang menjelaskan mengenai nilai indeks pencemaran serta status mutu pada setiap kelurahan. Terdapat beberapa data yang kurang relevan untuk dijadikan bahan penelitian.

3) DETERMINE DATA MINING GOAL

Penelitian dilakukan guna mengukur kinerja algoritma *naive bayes* untuk klasifikasi kualitas kondisi air tanah di DKI Jakarta menggunakan data 2019 dan 2021 dengan beberapa skenario yang berbeda. Dari hasil klasifikasi skenario tersebut, akan diketahui hasil algoritma *naive bayes* dengan kinerjanya. Selain itu, penelitian ini mempunyai tujuan untuk menemukan informasi baru terkait kualitas kondisi air tanah di DKI Jakarta dari data yang sudah diolah sehingga dapat menjadikan dasar atau bahan acuan untuk dilakukannya penelitian selanjutnya

B. DATA UNDERSTANDING

Pemahaman data dilakukan dengan menggunakan dua sumber data mengenai kualitas air tanah di DKI Jakarta, yaitu open data Jakarta dan website Dinas Lingkungan Hidup DKI Jakarta. Data yang digunakan dalam penelitian ini adalah data tahun 2019 dan 2021. Data untuk tahun 2019 diambil dari open data DKI Jakarta dalam bentuk file csv, sedangkan data untuk tahun 2021 diambil dari laporan pemantauan kualitas air

tanah yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta.

Alasan pemilihan kedua tahun tersebut adalah jarak waktu yang tidak terlalu jauh dari dilakukannya penelitian ini. Tahun 2020 tidak digunakan karena tidak dilakukannya pemantauan kualitas air tanah. Pada tahun 2020, hanya dilakukan penyusunan analisis kualitas air tanah dalam kurun waktu 5 tahun terakhir.

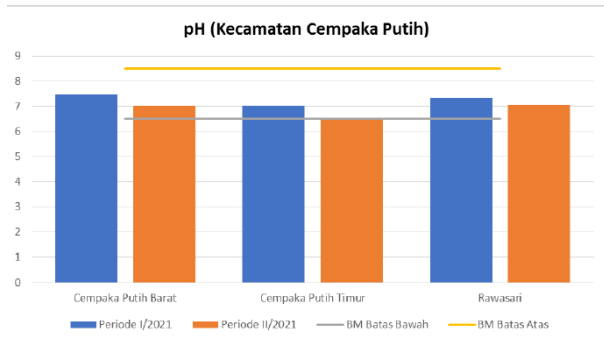
Sumber pertama adalah situs web open data DKI Jakarta pada tahun 2019 yang dikumpulkan secara terpisah antara periode 1 dan periode 2 dalam bentuk csv. Jumlah data kualitas air tanah yang terkumpul pada masing-masing periode adalah 267. Jumlah tersebut dihasilkan dari 5 Kota Administrasi di DKI Jakarta yaitu Jakarta Pusat, Jakarta Selatan, Jakarta Timur, Jakarta Utara dan Jakarta Barat. Parameter yang digunakan atas penilaian status mutu di data tahun 2019 diantaranya warna, kekeruhan, air raksa, besi, fluoride, cadmium, kesadahan, krom heksavalen, mangan, nitrat, nitrit, seng, sulfat, timah hitam, deterjen, zat organic, total coliform, dan e.coli. Gambar 3 merupakan dataset di tahun 2019

1	kecamatan,kelurahan,lintang_selatan,bujur_timur,parameter,hasil_analisa		
2	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Warna,21		
3	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Kekeruhan,0,98		
4	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Air Raksa,"<0,3"		
5	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Besi (Fe),"<0,8"		
6	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Fluorida,"0,43"		
7	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Cadmium (Cd),0		
8	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Kesadahan,"55,15"		
9	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Krom Heksavalen,"<0,3"		
10	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Mangan (Mn),"<0,5"		
11	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Nitrat,"0,22"		
12	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Nitrit,"0,6"		
13	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Seng (Zn),"0,1"		
14	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Sulfat,"40,85"		
15	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Timah Hitam (Pb),0		
16	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Detergen,"0,6"		
17	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Zat Organik,"1,82"		
18	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Total Coliform,100		
19	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,E.Coli,100		
20	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Indeks Pencemaran,"1,786,711,641"		
21	Gambir,Cideng,"06o10'05,1" " " " 106o48'33,5"" ,Kategori,Tercemar Ringan		

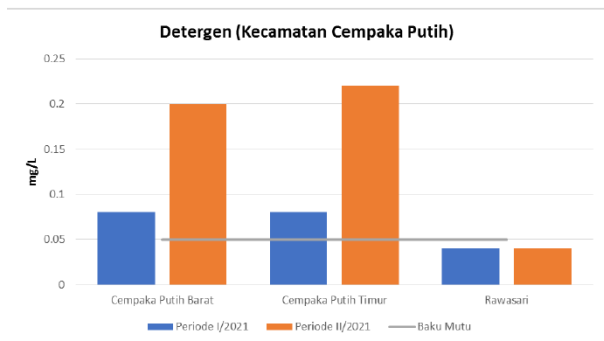
GAMBAR 3. Dataset 2019

Untuk sumber kedua yaitu didapatkan dari laporan kualitas air tanah di DKI Jakarta yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta dengan jumlah data dan Kota Adminstrasi yang digunakan sama dengan data di tahun 2019. Perbedaan dari kedua data tersbut terlak pada bentuk data yang didapatkan dan parameter yang digunakan guna klasifikasi status mutu air tanah. Bentuk data yang didapatkan dari laporan kualitas air tanah di DKI Jakarta yaitu grafik dan tabel dari 5 parameter yang mendominasi dan belum memenuhi -- baku mutu. Kelima parameter tersebut adalah pH, deterjen, mangan, e.coli, dan total coliform. Dengan contoh Kecamatan Cempaka Putih yang dibagi menjadi 3 kelurahan, Cempaka Putih Barat, Cempaka Putih Timur, dan Rawasari, Gambar 4 merupakan salah satu contoh bentuk data mengenai parameter pH, Gambar 5 merupakan salah satu contoh bentuk data mengenai parameter deterjen, Gambar 6 merupakan salah satu

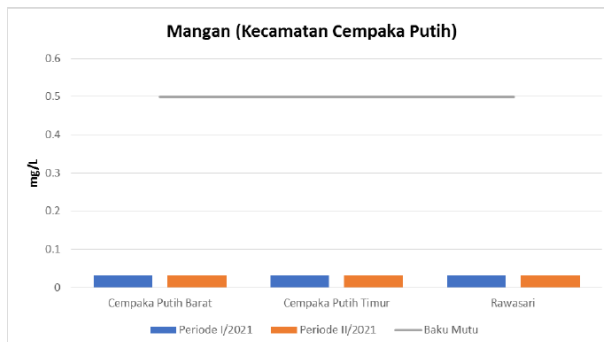
contoh bentuk data mengenai parameter mangan, Gambar 7 merupakan salah satu contoh bentuk data mengenai parameter total coliform, dan Gambar 8 merupakan salah satu contoh bentuk data mengenai parameter e.coli.



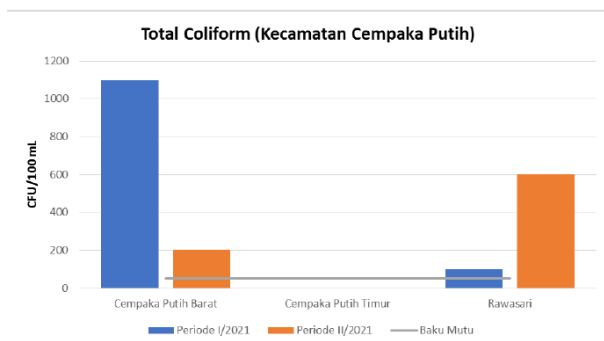
GAMBAR 4. Grafik pH Kecamatan Cempaka Putih



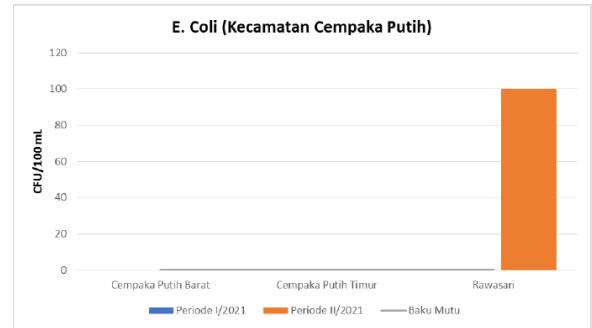
GAMBAR 5. Grafik Deterjen Kecamatan Cempaka Putih



GAMBAR 6. Grafik Mangan Kecamatan Cempaka Putih



GAMBAR 7. Grafik Total Coliform Kecamatan Cempaka Putih



GAMBAR 8. Grafik E.Coli Kecamatan Cempaka Putih

Selain paparan grafik dengan menggambarkan berapa nilai jumlah kandungan parameter pada tiap kelurahan pada 2 periode, terdapat paparan lanjutan mengenai hasil perhitungan kandungan 5 parameter yang menghasilkan nilai indeks pencemaran serta status mutu yang diperoleh. Tabel I adalah paparan dari hasil perhitungan grafik dari 5 parameter tersebut.

TABEL I
NILAI IP, STATUS MUTU, PARAMETER KRITIS KECAMATAN CEMPAKA PUTIH

Periode I				
No Titik	Kelurahan	Nilai IP	Status Mutu	Parameter Kritis
26	Cempaka Putih Barat	5.48	Cemar Sedang	Detergen Total Coliform
27	Cempaka Putih Timur	1.15	Cemar Ringan	Detergen
28	Rawasari	1.79	Cemar Ringan	Total Coliform
Periode II				
No Titik	Kelurahan	Nilai IP	Status Mutu	Parameter Kritis
26	Cempaka Putih Barat	2.87	Cemar Ringan	Detergen Total Coliform
27	Cempaka Putih Timur	2.99	Cemar Ringan	Detergen
28	Rawasari	4.54	Cemar Ringan	Total Coliform

Status mutu yang ditetapkan pada setiap kelurahan dinilai berdasarkan besarnya nilai IP atau indeks pencemaran. Terdapat kriteria yang telah dibuat oleh Dinas Lingkungan Hidup DKI Jakarta, yaitu, jika IP berada dalam rentang antara 0 hingga 1,0, maka kondisinya dapat diklasifikasikan sebagai "Baik," jika IP berada dalam rentang antara 1,0 hingga 5,0, maka kondisinya adalah "Cemar Ringan," ketika IP berada dalam rentang antara 5,0 hingga 10, kondisinya adalah "Cemar Sedang," terakhir, jika IP melebihi angka 10, maka kondisinya dapat diklasifikasikan sebagai "Cemar Berat".

Adanya perbedaan bentuk perolehan data, perlu dilakukannya perubahan atas data 2021 agar dapat mempermudah pembacaan, mempermudah melakukan sinkronisasi dengan data 2019, dan mempermudah

TABEL II
PERUBAHAN DATASET 2021

Kota	Kecamatan	Kelurahan	PH	Detergen	Mangan	Total Coliform	E_Coli	Indeks	Mutu
Jakarta Pusat	Gambir	Cideng	7,2	0,6	0,3	100	100	1,78	Cemar Ringan
Jakarta Pusat	Gambir	Duri Pulo	7,6	2,94	1,51	0	0	0,72	Baik
Jakarta Pusat	Gambir	Gambir	7,9	0,14	0,6	100	0	2,74	Cemar Ringan
Jakarta Pusat	Gambir	Kebon Kelapa	7,5	0,6	0,3	0	0	0,99	Baik
Jakarta Pusat	Gambir	Petojo Selatan	7,7	0,6	0,3	0	0	1,31	Cemar Ringan
Jakarta Pusat	Gambir	Petojo Utara	6,9	0,8	0,3	100	0	1,79	Cemar Ringan
Jakarta Pusat	Sawah Besar	Gunung Sahari Utara	7,5	0,2	0,29	0	0	0,44	Baik
Jakarta Pusat	Sawah Besar	Karang Anyar Utara	7,7	0,5	0,3	0	0	0,71	Baik
Jakarta Pusat	Sawah Besar	Kartini	7,9	0,4	0,3	0	0	0,58	Baik
Jakarta Pusat	Sawah Besar	Mangga Dua Selatan	7,3	0,18	0,35	100	0	2,7	Cemar Ringan
Jakarta Pusat	Cempaka Putih	Cempaka Putih Barat	7,3	0,9	0,8	0	0	1,62	Cemar Ringan
Jakarta Pusat	Cempaka Putih	Cempaka Putih Timur	7,3	0,5	0,3	0	0	0,71	Baik
Jakarta Pusat	Cempaka Putih	Rawasari	6,7	0,5	0,35	600	0	4,54	Cemar Ringan

preprocessing yang akan dilakukan pada penelitian ini. Tabel II merupakan hasil perubahannya dengan format dokumen di kecamatan Cempaka Putih serta beberapa kecamatan lainnya

C. DATA PREPARATION

Preprocessing dilakukan untuk menciptakan dataset klasifikasi kualitas air tanah di DKI Jakarta dapat digunakan dalam pemodelan naïve bayes. Beberapa tahapan ini merupakan cara agar mendapatkan hasil data yang bersih, terorganisir, dan lebih siap untuk dianalisa dalam penelitian ini.

1) PEMILIHAN ATRIBUT DAN LABEL

Berdasarkan *dataset* 2019, terdapat 6 atribut seperti yang dipaparkan pada Gambar 3. Sedangkan untuk *dataset* 2021, terdapat 10 atribut seperti yang dipaparkan pada Tabel II. Namun pada penelitian ini, atribut yang digunakan akan berdasarkan pada *dataset* di tahun 2021 karena data tersebut merupakan data terbaru yang diterbitkan oleh Dinas Lingkungan Hidup Jakarta. Oleh karena itu, Atribut yang akan digunakan berjumlah 10 diantaranya kota, kecamatan, kelurahan, pH, deterjen, mangan, total coliform, e.coli, indeks, dan status mutu. Berdasarkan atribut yang ada pada data, label yang akan digunakan pada penelitian ini adalah status mutu yang dibagi menjadi 4, yaitu Baik, Cemar Ringan, Cemar Sedang dan Cemar Berat.

2) NORMALISASI

Perolehan dataset dengan range nilai pada e.coli dan coliform yang terlampaui jauh dibutuhkannya normalisasi untuk meningkatkan *accuracy* model. Dihasilkan normalisasi dengan nilai pada setiap atribut numerik dari rentang 0 - 1.

D. MODELLING

Pemodelan dilakukan menggunakan algoritma *naïve bayes* dengan 2 teknik, yaitu *cross validation* serta *percentage split*.

Teknik *cross validation* dilakukan dengan 6 skenario yang terdiri dari nilai *fold* 5 sampai dengan *fold* 10.

Teknik *percentage split* dilakukan dengan 4 skenario yang terdiri dengan perbandingan rasio 80 % : 20%, 70 % : 30%, 60 % : 50%, dan 50 % : 50%. Rasio pemisahan terdiri *data training* dan *data testing*

TABEL III
ACCURACY DAN ERROR RATE

Cross Validation		
Scenario	Accuracy	Error Rate
Skema 1	83,89%	16,10%
Skema 2	83,33%	16,66%
Skema 3	83,33%	16,66%
Skema 4	83,80%	16,19%
Skema 5	83,33%	16,66%
Skema 6	83,80%	16,19%
Percentage split		
Scenario	Accuracy	Error Rate
Skema 1	84,36%	15,63%
Skema 2	84,26%	15,73%
Skema 3	83,61%	16,38%
Skema 4	83,23%	16,76%

TABEL IV
PRECISION, RECALL, DAN F1-SCORE

Cross Validation			
Scenario	Precision	Recall	F1-Score
Skema 1	0,8478	0,839	0,8389
Skema 2	0,843	0,8333	0,8333
Skema 3	0,8431	0,8333	0,8332
Skema 4	0,8488	0,838	0,8382
Skema 5	0,844	0,8333	0,8333
Skema 6	0,8499	0,838	0,8379
Percentage split			
Scenario	Precision	Recall	F1-Score
Skema 1	0,8566	0,8436	0,8436
Skema 2	0,8556	0,8427	0,8426
Skema 3	0,8517	0,8361	0,8361
Skema 4	0,8478	0,8324	0,8322

E. EVALUATION

Berdasarkan 10 skema yang sudah dipaparkan pada tahap *modelling*, dengan 6 skema *cross validation* dan 4 skema *percentage split*, Tabel III adalah hasil *accuracy*, *error rate*, untuk Tabel IV adalah hasil *precision*, *recall*, dan *f1-score* dari 10 skema yang dilakukan.

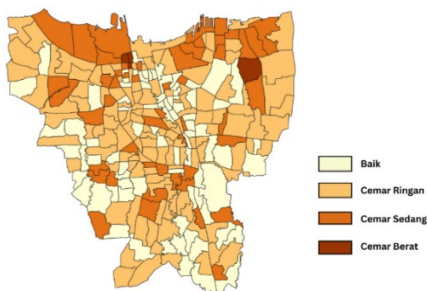
Pengujian *cross validation* pada skema 1 dengan folds 5 menghasilkan akurasi tertinggi yaitu 83.89 % dengan *error rate* 16.10% dan juga menempati nilai tertinggi pada *precision* dengan 0.8478, *recall* 0.8390, dan *F1-score* 0.8389, seperti yang dipaparkan pada Tabel III dan Tabel IV.

Pengujian *percentage split* pada skema 1 dengan rasio 80:20 menghasilkan nilai *accuracy* tertinggi yaitu 84.36% dengan *error rate* 15.63%. dan juga menempati nilai tertinggi pada *precision* 0.8566, *recall* 0.8436, dan *F1-score* 0.8436.

F. DEPLOYMENT

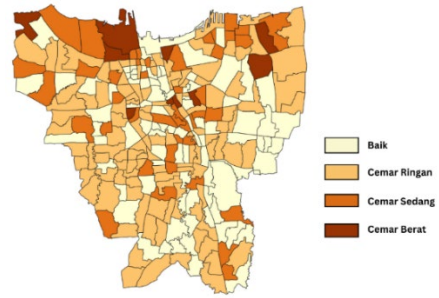
Deployment pada penelitian klasifikasi kualitas dan prediksi kondisi air tanah di DKI Jakarta dengan menggunakan algoritma naïve bayes, menghasilkan laporan penelitian. Laporan akhir tersebut berisi paparan tentang penggunaan teknik *modelling cross validation* dan *percentage split* pada data kualitas air tanah 2019 dan 2021 yang telah melalui tahap *preprocessing* sehingga diperoleh data yang bersih dan terorganisir. Selanjutnya, dilakukan perbandingan *accuracy* serta kinerja teknik algoritma terbaik yang dapat dijadikan dasar atau bahan acuan untuk penelitian selanjutnya.

Selain pembuatan laporan akhir, terdapat paparan mengenai pemantauan kualitas air tanah di DKI Jakarta dengan 2 periode di tahun 2019 serta 2 periode di tahun 2021.



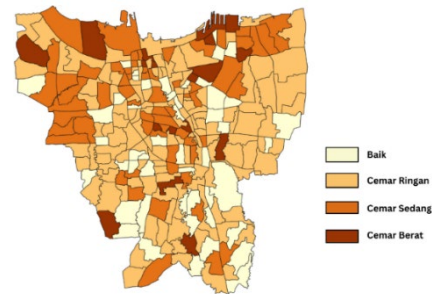
GAMBAR 9. Pemetaan Kualitas Air Tanah Tahun 2019 Periode 1

Berdasarkan data yang didapatkan dari Open Data DKI Jakarta dan visualisasi pada Gambar 9, untuk kualitas air tanah di DKI Jakarta pada tahun 2019 periode 1 terdapat 4 wilayah dengan status mutu air tanah cemar berat, diantaranya yaitu Kota Administrasi Jakarta Utara pada kelurahan Sukapura, dan Kota Administrasi Jakarta Barat pada kelurahan Pekojan, Roamalaka, dan Tambora. Selain adanya kelurahan yang memiliki status mutu air cemar berat, pada Kota Administrasi Jakarta Utara mendominasi memiliki status mutu air cemar sedang.



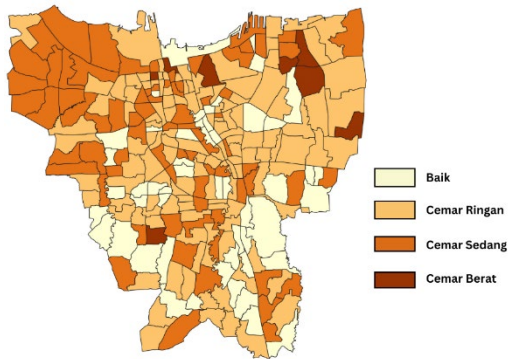
GAMBAR 10. Pemetaan Kualitas Air Tanah Tahun 2019 Periode 2

Berdasarkan data yang didapatkan dari Open Data DKI Jakarta dan visualisasi pada Gambar 10, untuk kualitas air tanah di DKI Jakarta pada tahun 2019 periode 2 terdapat kenaikan jumlah wilayah dengan status mutu air tanah Cemar Berat dengan total 10 kelurahan, diantaranya yaitu Kota Administrasi Jakarta Utara pada kelurahan Pademangan Barat, Pejagalan, Pluit, Semper Timur, dan Sukapura. Kota Administrasi Jakarta Pusat pada kelurahan Cempaka Putih Barat, Kramat dan Kwitang. Kota Administrasi Jakarta Barat pada kelurahan Kamal dan Slipi. Untuk dominasi status mutu air cemar sedang berada pada Kota Administrasi Jakarta Barat.



GAMBAR 11. Pemetaan Kualitas Air Tanah Tahun 2021 Periode 1

Berdasarkan data yang didapatkan dari laporan akhir pemantauan kualitas air tanah tahun 2021 yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta dan visualisasi pada Gambar 11, untuk kualitas air tanah di DKI Jakarta pada tahun 2021 periode 1 terdapat kenaikan jumlah wilayah dengan status mutu air tanah Cemar Berat dengan total 18 kelurahan, diantaranya yaitu Kota Administrasi Jakarta Utara pada kelurahan Kapuk Muara, Koja, Sunter Jaya, Tanjung Priok, Tugu Selatan, dan Warakas. Kota Administrasi Jakarta Pusat pada kelurahan Serdang. Kota Administrasi Jakarta Barat pada kelurahan, Glodok, Pegadungan, Pinangsia dan Tangki. Kota Administrasi Jakarta Selatan pada kelurahan Kalibata, Lebak Bulus, Pasar Manggis, dan Setiabudi. Kota Administrasi Jakarta Timur pada kelurahan Cijantung, Cipinang Muara dan Kebon Manggis.



GAMBAR 12. Pemetaan Kualitas Air Tanah Tahun 2021 Periode 2

Berdasarkan data yang didapatkan dari laporan akhir pemantauan kualitas air tanah tahun 2021 yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta dan visualisasi pada Gambar 12, untuk kualitas air tanah di DKI Jakarta pada tahun 2021 periode 2 terdapat penurunan jumlah wilayah dengan status mutu air tanah Cemar Berat dengan total 9 kelurahan, diantaranya yaitu Kota Administrasi Jakarta Utara pada kelurahan Pademangan Timur, Semper Barat, Sukapura, dan Tugu Selatan. Kota Administrasi Jakarta Barat pada kelurahan, Jembatan Lima dan Pinangsia. Kota Administrasi Jakarta Selatan pada kelurahan Cipete Selatan. Kota Administrasi Jakarta Timur pada kelurahan Ujung Menteng. Untuk dominasi status mutu air cemar sedang berada pada Kota Administrasi Jakarta Barat.

V. KESIMPULAN

Hasil evaluasi algoritma naïve bayes dalam klasifikasi kualitas kondisi air tanah di DKI Jakarta berdasarkan dua teknik, yaitu *cross-validation* dan *percentage split*, menghasilkan nilai akurasi yang berbeda. Akurasi tertinggi tercapai dengan menggunakan teknik *percentage split* 80:20, mencapai 84.36%, dengan nilai presisi 0.8566, recall 0.8436, dan F1-score 0.8436.

Selain itu, hasil pemantauan kualitas air tanah di DKI Jakarta pada tahun 2019 dan 2021, dengan dua periode pada setiap tahunnya, menggambarkan tren tertentu. Pada tahun 2019, periode 2 mengalami peningkatan jumlah pencemaran berat, dan tren serupa terlihat pada tahun 2021, periode 1. Namun, pada tahun 2021, periode 2 mengalami peningkatan pencemaran sedang, sementara terjadi penurunan pencemaran berat. Pencemaran ini terjadi akibat terkaitnya dengan lima parameter yang melebihi baku mutu, yaitu pH, deterjen, mangan, E. coli, dan total coliform.

Parameter tersebut dapat menyebabkan masalah kesehatan yang signifikan. Misalnya, ketika tingkat pH air tanah terlalu tinggi atau terlalu rendah, menurut *National Institutes of Health* ini dapat memiliki dampak serius pada tubuh manusia. Tingkat pH yang tinggi disebut "basa" atau "alkalis." Jika tingkat pH darah kita terlalu tinggi, hal ini dapat menyebabkan otot berkedut, mual, kebingungan, koma, dan dampak negatif kesehatan lainnya. Selain itu, kelebihan mangan dalam air tanah juga dapat memiliki dampak kesehatan yang serius.

Gejala keracunan mangan meliputi gemetar, kejang otot, masalah pendengaran, mania, insomnia, depresi, kehilangan nafsu makan, sakit kepala, mudah tersinggung, lemah, dan perubahan suasana hati. Oleh karena itu, penting untuk memantau dan mengendalikan parameter-parameter ini guna memastikan kualitas air tanah yang baik dan mencegah dampak negatif pada kesehatan manusia.

PERAN PENULIS

Arsya Fathiarahma: Konseptual, Persiapan Data, Penyusunan Draft Asli;

Nina Sulistiyowati: Validasi, Penulisan Penyusunan Draft Asli, Penulisan Review & Penyuntingan;

Taufik Ridwan: Penulisan Draft Asli Persiapan, Penulisan Review & Editing;

Apriade Voutama: Visualisasi, Penulisan Review & Penyuntingan;

COPYRIGHT



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

DAFTAR PUSTAKA

- [1] Detik, "Ada Berapa Banyak Jumlah Air di Bumi? Ini Jawabannya," 2022. [Online]. Available: [https://www.detik.com/edu/detikpedia/d-5880228/ada-berapa-banyak-jumlah-air-di-bumi-ini-jawabannya#:~:text=Merangkul laman Lembaga Survei Geologi,juta kilometer kubik \(km3\)](https://www.detik.com/edu/detikpedia/d-5880228/ada-berapa-banyak-jumlah-air-di-bumi-ini-jawabannya#:~:text=Merangkul%20laman%20Lembaga%20Survei%20Geologi,juta%20kilometer%20kubik%20(km3).).
- [2] H. Hamzar, S. Suprpta, and A. Amal, "Analisis Kualitas Air Tanah Dangkal Untuk Keperluan Air Minum Di Kelurahan Bontononpo Kecamatan Bontononpo Kabupaten Gowa," *J. Environ. Sci.*, vol. 3, no. 2, 2021, doi: 10.35580/jes.v3i2.20048.
- [3] A. Misa, R. S. Duka, S. Layuk, and Y. T. Kawatu, "HUBUNGAN KEDALAMAN SUMUR BOR DENGAN KADAR BESI (Fe) DAN MANGAN (Mn) DI KELURAHAN MALENDENG KECAMATAN PAAL 2 KOTA MANADO," *JKL*, vol. 9, no. 5, p. 55, 2019.
- [4] PDAM Tirta benteng, "Amankah Mengonsumsi Air Tanah, Ini Penjelasan Ahli Hidrologi," 2018. <http://www.pdamtirtabenteng.co.id/berita/amankah-mengonsumsi-air-tanah-ini-penjelasan-ahli-hidrologi>
- [5] M. J. Sodiq and E. I. Sela, "Perbandingan Metode Naive Bayes Dan K-Nearest Neighbor Pada Klasifikasi Kualitas Udara Di Dki Jakarta," 2019.
- [6] D. Purwanto, "Klasifikasi Naive Bayes Untuk Memprediksi Kualitas Udara Berdasarkan Indeks Standar Pencemar Udara (Ispu)," pp. 1–94, 2019, [Online]. Available: <https://123dok.com/document/y8rk480q-skripsi-klasifikasi-memprediksi-kualitas-berdasarkan-indeks-standar-pencemar.html>
- [7] J. H. Jaman, J. H. Jaman, . C., and A. R. Sanjaya, "Klasifikasi jenis mobil paling diminati Di Indonesia menggunakan algoritma Naive bayes," *Fakt. Exacta*, vol. 13, no. 1, p. 18, 2020, doi: 10.30998/faktorexacta.v13i1.5573.
- [8] A. Aziiz, H. Kirono, I. Asror, Y. Firdaus, and A. Wibowo, "Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma Naive Bayes," *eProceedings ...*, vol. 9, no. 3, pp. 1962–1969, 2022, [Online]. Available: <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18002%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18002/17631>
- [9] F. Husna, R. Ramadhan, A. C. Firdaus, and I. Veritawati,

- “Prediksi Pembatalan Reservasi Hotel Menggunakan Algoritma Naive Bayes,” vol. 4, no. 1, pp. 76–80, 2023.
- [10] A. Yoga Pratama *et al.*, “Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja),” *J. Sains Komput. Inform. (J-SAKTI)*, vol. 5, no. 2, pp. 897–910, 2021.
- [11] L. Muflikhah, D. E. Ratnawati, and R. R. MP, *Data Mining*. 2018. [Online]. Available: https://books.google.co.id/books?id=V_NqDwAAQBAJ&printec=copyright&redir_esc=y#v=onepage&q&f=false
- [12] A. Amrullah, I. Purnamasar, B. N. Sar, Garno, and A. Voutama, “Analisis Cluster Faktor Penunjang Pendidikan Menggunakan Algoritma K-Means (Studi Kasus : Kabupaten Karawang),” *JIRE (Jurnal Inform. Rekayasa Elektron.)*, vol. 5, no. 2, 2022.
- [13] M. R. Nahjan, N. Heryana, and A. Voutama, “Implementasi Rapidminer Dengan Metode Clustering K-Means untuk Analisa Penjualan pada Toko Oj Cell,” *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 1, 2023.
- [14] H. Nuraliza, O. N. Pratiwi, and F. Hamami, “Analisis Sentimen IMDb Film Review Dataset Menggunakan Support Vector Machine (SVM) dan Seleksi Feature Importance,” *J. Mirai Manaj.*, vol. 7, no. 1, pp. 1–17, 2022.
- [15] M. F. Andriansyah, D. Yusup, and A. Voutama, “Menggunakan Metode Naïve Bayes Berbasis Website Web-Based Expert System of Covid-19 Early Detection Using Naïve Bayes Method,” *J. Inf. Technol. Comput. Sci.*, vol. 4, no. 2, pp. 446–455, 2021.
- [16] N. M. Maghfur, F. Muhammad, and A. Voutama, “Analysis of the Relationship between Public Sentiment on Social Media and Indonesian Covid-19 Dynamics,” vol. 3, no. 3, pp. 336–345, 2021.
- [17] F. V. Sari and A. Wibowo, “Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi,” *J. SIMETRIS*, vol. 10, no. 2, pp. 681–686, 2019.
- [18] F. Tempola, M. Muhammad, and A. Khairan, “Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.
- [19] S. Saadah and H. Salsabila, “Prediksi Harga Bitcoin Menggunakan Metode Random Forest,” *J. Politek. Caltex Riau*, vol. 7, no. 1, pp. 24–32, 2021.
- [20] A. Primajaya, B. N. Sari, and A. Khusaeri, “Prediksi Potensi Kebakaran Hutan dengan Algoritma Klasifikasi C4.5 Studi Kasus Provinsi Kalimantan Barat,” *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 188, 2020, doi: 10.26418/jp.v6i2.37834.
- [21] Y. Rizki, R. Medikawati Taufiq, H. Mukhtar, and D. Putri, “Klasifikasi Pola Kain Tenun Melayu Menggunakan Faster R-CNN,” *IT J. Res. Dev.*, vol. 5, no. 2, pp. 215–225, 2021, doi: 10.25299/itjrd.2021.vol5(2).5831.