

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Air merupakan kebutuhan pokok yang tidak dapat dipisahkan dari kehidupan manusia, baik untuk kebutuhan rumah tangga maupun industri. Air dapat diperoleh dari berbagai sumber seperti air tanah, air sungai, ataupun sistem air terpusat. Di Provinsi DKI Jakarta, 35% dari penduduk DKI Jakarta menggunakan air tanah untuk pemenuhan air baku (Yonatan et al., 2021). Untuk memastikan keamanan air yang digunakan, air harus memenuhi standar kualitas yang ketat dan terkontrol dengan baik. Kementerian Kesehatan RI menetapkan bahwa air bersih harus memenuhi parameter baku mutu yang tertuang dalam Permenkes No. 2 tahun 2023. Kemudian data yang telah dipantau akan dilakukan pengukuran untuk menentukan tingkat pencemaran berdasarkan peraturan yang diatur dalam Permen LHK No. 37 tahun 2021. Dengan melakukan pemantauan dan pengukuran secara rutin dapat memastikan air tersebut layak digunakan oleh masyarakat.

Seiring berkembangnya teknologi, pengukuran tingkat pencemaran pada kualitas air dapat dilakukan dengan mudah. Dengan memanfaatkan *machine learning*, parameter atau variabel hasil pemantauan akan digunakan dalam mengetahui kualitas air dengan metode klasifikasi data (Mutoffar & Fadillah, 2022). *Machine Learning* merupakan sistem yang dirancang dengan kemampuan adaptif dalam meningkatkan kinerja dan efisiensinya (Heryadi & Wahyono, 2020). Dalam klasifikasi data, *machine learning* bermanfaat dalam beradaptasi untuk meningkatkan kinerja dalam melakukan klasifikasi variabel target.

Dalam pelaksanaannya, metode klasifikasi dapat menggunakan beberapa model seperti *Naive Bayes*, *Random Forest*, KNN dan masih banyak lagi. Penerapan model klasifikasi pada kualitas air di DKI Jakarta ini sudah pernah dilakukan pada penelitian sebelumnya. Pada penelitian Mutoffar dan Fadillah (2022), dilakukan klasifikasi berdasarkan data kualitas air sumur tahun 2017. Dengan menggunakan model *Random Forest*, didapatkan hasil presisi 82,3 % dan sensitivitas 83%. Hasil ini diperoleh dengan membagi data latih dan data uji

dengan proposi 80:20, serta mengatur parameter percabangan sampel minimum adalah 5 dan jumlah pohon keputusan adalah 10. Kemudian penelitian serupa pernah dilakukan juga oleh Fathiarahma dkk (2023), dimana data yang digunakan merupakan data pemantauan tahun 2019 dan 2020. Dengan menggunakan model klasifikasi *Naive Bayes* dengan pembagian data latih dan data uji sebesar 80:20, diperoleh nilai akurasi 84,36%.

Akan tetapi, hasil yang di dapatkan pada penelitian tersebut belum maksimal, salah satunya terdapat proses pengolahan data yang terlewat. Pada penelitian Mutoffar dan Fadillah (2022), tahapan pengolahan data tidak melalui proses pembersihan data (*data cleansing*) dimana hal ini penting untuk menentukan apakah data memiliki data yang kosong, data duplikat, ataupun adanya pencilan (*outlier*). Kemudian tahapan penelitian ini tidak mempertimbangkan untuk melihat korelasi antar variabel serta melakukan seleksi fitur (*feature selection*), langkah ini berfungsi untuk melihat dan menentukan fitur apa saja yang akan dipakai dalam pemodelan. Pada langkah selanjutnya data tidak dipertimbangkan untuk dilakukan normalisasi data yang bermanfaat dalam mengatasi *skewness* pada persebaran data, mengurangi dampak pencilan (*outlier*), serta meningkatkan kinerja model.

Kemudian, pada penelitian yang dilakukan oleh Fathiarahma dkk (2023) telah dilakukan dengan lebih baik. Akan tetapi, masih belum mempertimbangkan untuk dilakukan proses pembersihan data (*data cleansing*), serta tidak mempertimbangan untuk melihat korelasi variabel dan seleksi fitur (*feature selection*).

Selain itu, terdapat permasalahan lain dari data yang digunakan. Hal ini dikarenakan pada penelitian terdahulu merujuk pada Permenkes no. 32 tahun 2017, dimana regulasi tersebut telah digantikan oleh Permenkes no. 2 tahun 2023. Terdapat perbedaan variabel pemantauan serta pengetatan baku mutu pada regulasi terbaru, seperti yang ditampilkan pada Tabel 1.1 berikut.

Tabel 1.1 Perbedaan Variabel dan Baku Mutu Kualitas Air

No.	Jenis Parameter	Kadar Maksimum yang Diperbolehkan		Satuan
		Permenkes 32/2017	Permenkes 2/2023	
Mikrobiologi				
1	Escherichia coli	0	0	CFU/100ml
2	Total Coliform	50	0	CFU/100ml
Fisik				
3	Suhu	Suhu udara \pm 3	Suhu udara \pm 3	$^{\circ}$ C
4	Total Dissolve Solid	1000	300	mg/L
5	Kekeruhan	25	3	NTU
6	Warna	50	10	TCU
7	Bau	Tidak Berbau	Tidak Berbau	-
8	Rasa	Tidak Berasa	(tidak diatur)	-
Kimia				
9	pH	6.5 – 8.5	6.5 – 8.5	-
10	Nitrat (sebagai NO ³) (terlarut)	10	20	mg/L
11	Nitrit (sebagai NO ²) (terlarut)	1	3	mg/L
12	Kromium valensi 6 (Cr ⁶⁺) (terlarut)	0.05	0.01	mg/L
13	Besi (Fe) (terlarut)	1	0.2	mg/L
14	Mangan (Mn) (terlarut)	0.5	0.1	mg/L
15	Sisa khlor (terlarut)	(tidak diatur)	0.2 – 0.5 dengan waktu kontak 30 menit	mg/L
16	Arsen (As) (terlarut)	0.05	0.01	mg/L
17	Kadmium (Fb) (terlarut)	0.005	0.003	mg/L
18	Timbal (Pb) (terlarut)	0.05	0.01	mg/L
19	Fluoride (F) (terlarut)	1.5	1.5	mg/L
20	Alumunium (Al) (terlarut)	(tidak diatur)	0.2	mg/L
21	Kesadahan (CaCO ₃)	500	(tidak diatur)	mg/L
22	Sianida	0.1	(tidak diatur)	mg/L
23	Deterjen	0.05	(tidak diatur)	mg/L
24	Pestisida Total	0.1	(tidak diatur)	mg/L
25	Air Raksa	0.001	(tidak diatur)	mg/L
26	Selenium	0.01	(tidak diatur)	mg/L
27	Seng	15	(tidak diatur)	mg/L
28	Sulfat	400	(tidak diatur)	mg/L

Sumber: Kementerian Kesehatan

Regulasi terbaru ini sudah diterapkan mulai tahun pencatatan 2023, sementara ditahun pencatatan sebelumnya masih menggunakan regulasi lama. Dengan perbedaan regulasi ini akan mempengaruhi hasil dari status pencemaran air yang juga berimbas pada variabel target yang digunakan dalam model klasifikasi, sehingga perlu dilakukan pemodelan ulang sesuai dengan menggunakan data terbaru yang sesuai dengan regulasi yang berlaku.

Terdapat banyak model klasifikasi yang dapat digunakan dalam *machine learning*. Dengan mengacu pada penelitian serupa, bila diterapkan untuk mengklasifikasi kualitas udara, maka model *K-Nearest Neighbors* merupakan model yang lebih baik untuk digunakan dibandingkan dengan *Naive Bayes* dan *Random Forest*. Dalam penelitian yang dilakukan oleh Amalia dkk (2022), diperoleh akurasi tertinggi sebesar 96% dengan menggunakan model KNN. Sementara itu pada penelitian yang dilakukan oleh Nugroho (2023), diperoleh nilai akurasi tertinggi sebesar 90% dengan menggunakan model *Random Forest* dan pada penelitian yang dilakukan oleh Kirono (2022), diperoleh akurasi tertinggi sebesar 88% dengan menggunakan *Naive Bayes*. Bila melihat potensi yang dihasilkan, metode KNN berpotensi menghasilkan performa yang lebih baik dari penelitian yang telah dilakukan sebelumnya.

Permasalahan selanjutnya, penelitian sebelumnya dilaksanakan hingga dilakukan pemodelan tanpa adanya *interpretation* yang diterapkan. Bila mengacu pada metode KDD (*Knowledge Discovery in Database*), *Interpretation* merupakan tahapan metodologi yang mengimplementasikan hasil penerapan ke dalam *final report* salah satunya dengan diterapkan dalam bentuk visualisasi data. Visualisasi data berperan dalam mengubah data yang kompleks menjadi bentuk yang lebih mudah dimengeti (Enterprise, 2024). Selain itu, Dinas Lingkungan Hidup DKI Jakarta telah melaporkan dan mempublikasikan secara publik ringkasan pemantauan kualitas air di DKI Jakarta tahun 2023. Ringkasan hasil pemantauan yang dipublikasi merupakan data berbentuk tabel, hal ini akan menyulitkan dalam melakukan pelacakan data (Machda & Machda, 2013).

Dari pembahasan diatas dapat diketahui bahwa masih adanya sudah adanya penelitian terkait penerapan teknologi *machine learning* dalam klasifikasi air sumur di DKI Jakarta. Akan tetapi data yang sudah tidak relevan dengan regulasi terbaru. Kemudian adanya potensi untuk menggunakan KNN sebagai model klasifikasi dapat meningkatkan performa dari *machine learning*. Selain itu, dengan belum tuntasnya *interpretation* dan adanya potensi sulitnya melakukan pelacakan data kualitas air, maka visualisasi data dapat berperan sebagai salah satu *interpretation* dari hasil penelitian yang telah dilakukan serta dapat

memudahkan dalam pelacakan data. Sehingga penelitian ini bertujuan untuk menerapkan model klasifikasi pada tahun 2023 menggunakan metode KNN, serta melakukan *interpretation* pada hasil klasifikasi sehingga diperoleh performa *machine learning* yang lebih baik dari penelitian sebelumnya dan membantu dalam pelacakan data kualitas air.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dijabarkan, maka dapat diidentifikasi beberapa masalah sebagai berikut:

1. Penerapan *machine learning* yang belum maksimal dan data yang sudah tidak relevan dalam melakukan klasifikasi air sumur di DKI Jakarta
2. Belum adanya penelitian klasifikasi kualitas air sumur di DKI Jakarta 2023
3. Belum adanya penelitian klasifikasi kualitas air sumur di DKI Jakarta menggunakan model *K-Nearest Neighbors*
4. Belum adanya visualisasi data (*dashboard*) untuk data kualitas air sumur di DKI Jakarta

1.3 Pembatasan Masalah

Penelitian mengenai klasifikasi air sumur memiliki sejumlah batasan yang perlu dipertimbangkan, yaitu

1. Penelitian ini akan berfokus pada penerapan model klasifikasi dan visualisasi data pada kualitas air sumur di Provinsi DKI Jakarta
2. Data yang dapat digunakan adalah Data Kualitas Air Sumur 2023 yang diterbitkan oleh Dinas Lingkungan Hidup DKI Jakarta
3. Baku mutu air yang digunakan mengacu Permenkes no. 2 tahun 2023
4. Penelitian ini akan menerapkan model klasifikasi *K-Nearest Neighbors*
5. Pembuatan Visualisasi Data (*Dashboard*) menggunakan *tools* dari aplikasi Looker Studio

1.4 Perumusan Masalah

Berdasarkan latar belakang dan identifikasi masalah di atas, maka perumusan masalah penelitian ini adalah:

1. Bagaimana penerapan model klasifikasi *machine learning* dalam memprediksi kualitas air sumur di DKI Jakarta tahun 2023?
2. Bagaimana penerapan visualisasi data pada kualitas air sumur di DKI Jakarta tahun 2023?

1.5 Tujuan Penelitian

Tujuan penelitian ini adalah untuk:

1. Melakukan seleksi data kualitas air sumur di DKI Jakarta tahun 2023 yang akan digunakan dalam melakukan pemodelan
2. Membagi data kualitas air sumur di DKI Jakarta tahun 2023 menjadi data latih dan data uji
3. Melatih model klasifikasi pada data latih kualitas air sumur di DKI Jakarta tahun 2023
4. Menguji dan mengetahui performa penerapan model klasifikasi pada data uji kualitas air sumur di DKI Jakarta tahun 2023
5. Membandingkan hasil prediksi (*prediction class*) dengan hasil aktual (*actual class*) dari pengujian model pada data uji kualitas air sumur di DKI Jakarta 2023
6. Mengimplementasikan hasil pemodelan kedalam visualisasi data

1.6 Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

1. Referensi untuk menambah pengetahuan tentang klasifikasi kualitas air sumur dan penerapan algoritma *K-Nearest Neighbors* sebagai salah satu metode untuk penelitian selanjutnya
2. Dapat memudahkan dalam melakukan pelacakan data pada kualitas air sumur di DKI Jakarta tahun 2023
3. Dapat berkontribusi dalam pengembangan akademis tentang penerapan *machine learning* di bidang lingkungan hidup