BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka dilakukan untuk memperdalam pemahaman terhadap konsep, metode, serta hasil-hasil penelitian terdahulu. Kegiatan ini mengacu pada berbagai sumber literatur guna membangun dasar ilmiah yang kuat dalam menganalisis dan memecahkan permasalahan utama dalam penelitian.

2.1.1 Kualitas Air Sumur

Air merupakan kebutuhan esensial yang sangat penting bagi kelangsungan hidup manusia dan makhluk hidup lainnya. Namun, di kawasan dengan kepadatan penduduk tinggi seperti Jakarta Timur, kualitas air sumur cenderung menurun akibat pencemaran lingkungan. Berdasarkan Peraturan Menteri Kesehatan Nomor 32 Tahun 2017, kualitas air harus diawasi melalui sejumlah parameter fisik, kimia, dan biologi, termasuk pH, TDS, kandungan Besi (Fe), Nitrat sebagai N, E. Coli, Total Coliform, warna, dan bau, guna memastikan bahwa air yang dikonsumsi masyarakat tetap aman dan memenuhi standar kesehatan. (Permenkes No. 32, 2017).

2.1.2 Data Mining

Data mining adalah proses memperoleh informasi bermakna dari kumpulan data besar dengan memanfaatkan teknik-teknik matematika, statistik, dan kecerdasan buatan. Hasil ekstraksi ini dapat dimanfaatkan untuk mendukung pengambilan keputusan dan melakukan prediksi pada berbagai bidang, seperti

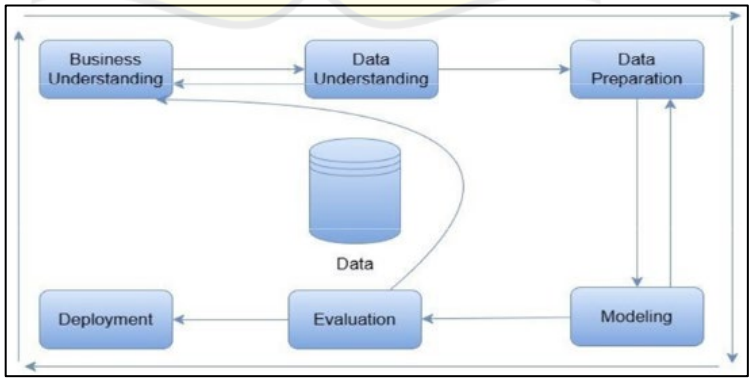
bisnis, penelitian ilmiah, teknologi, kesehatan, dan sebagainya. (Muttaqin et al., 2023).

2.1.3 Klasifikasi

Klasifikasi adalah proses untuk mengetahui nilai suatu data dan menetapkan data tersebut ke dalam salah satu kelas yang telah ditentukan sebelumnya. Proses ini melibatkan pembangunan model berdasarkan data latih yang tersedia, yang kemudian digunakan untuk mengelompokkan atau memprediksi kelas pada data baru. (Mutoffar & Fadillah, 2022).

2.1.4 CRISP-DM sebagai Tahap Merancang Pengembangan Sistem

CRISP-DM (Cross Industry Standard Process for Data Mining) adalah metode yang umum digunakan oleh para ahli dalam proses pemodelan data. Tujuan utama dari metode ini adalah untuk menemukan pola-pola yang relevan dan bermakna dari data yang dianalisis. CRISP-DM menawarkan kerangka kerja yang sistematis dan terstruktur, sehingga memudahkan pengguna untuk mengikuti langkah-langkah yang jelas dalam proses penelitian. Metode ini terdiri dari enam tahapan, yaitu:



Gambar 2. 1 Alur Metodologi CRISP-DM (Fitrianti et al., 2023)

Dimana berdasarkan Gambar 2.1 di atas, Adapun penjelasan dari setiap alur CRISP-DM yaitu:

1. Business Understanding

   Business Understanding merupakan tahap awal dalam proses data mining yang bertujuan untuk menetapkan sasaran bisnis, memahami kondisi dan konteks yang melatarbelakangi penelitian, serta merumuskan tujuan penelitian ke dalam bentuk permasalahan yang dapat diselesaikan melalui pendekatan data mining. (Dhewayani et al., 2022).

2. Data Understanding

   Data Understanding merupakan tahap awal dalam pengolahan data yang mencakup proses pengumpulan data awal, pemeriksaan data yang akan digunakan, serta evaluasi terhadap kualitas data tersebut. Pada tahap ini, setiap fitur dalam data akan dideskripsikan untuk memberikan pemahaman lebih lanjut mengenai karakteristik data secara menyeluruh. (Dhewayani et al., 2022).

3. Data Preparation

   Data Preparation adalah tahap yang dilakukan setelah data berhasil dikumpulkan. Pada proses ini, data akan melalui serangkaian langkah seperti identifikasi, pemilihan data yang relevan, pembersihan dari nilai yang tidak sesuai atau hilang, serta transformasi data agar siap digunakan dalam tahap pemodelan selanjutnya. (Dhewayani et al., 2022).

4. Modeling

   Modeling adalah tahap di mana algoritma mulai diterapkan untuk mencari, mengidentifikasi, dan membentuk pola dari data yang telah disiapkan. Pola-pola ini nantinya akan digunakan dalam analisis data penelitian guna mendukung proses klasifikasi atau prediksi. (Dhewayani et al., 2022).

5. Evaluation

   Evaluation merupakan proses untuk menilai hasil dari model yang telah dibangun pada tahap pemodelan. Tahap ini bertujuan untuk mengukur kinerja model yang dihasilkan, serta meninjau sejauh mana model tersebut merepresentasikan proses data mining yang telah dilakukan. Hasil evaluasi ini juga membantu dalam menentukan model terbaik yang layak digunakan untuk analisis lebih lanjut. (Dhewayani et al., 2022).

6. Deployment

   Deployment merupakan tahap akhir dalam proses data mining yang melibatkan penyusunan laporan atau publikasi ilmiah, seperti artikel jurnal, berdasarkan hasil dari penelitian yang telah dilakukan. Tahap ini bertujuan untuk menyampaikan temuan dan insight yang diperoleh kepada pihak terkait atau khalayak yang lebih luas. (Dhewayani et al., 2022).

2.1.5 Algoritma *K-Nearest Neighbor* (KNN)

Algoritma K-Nearest Neighbor (KNN) merupakan salah satu metode supervised learning yang banyak digunakan dalam tugas klasifikasi dan pengenalan pola. Pada metode ini, suatu data baru diklasifikasikan berdasarkan mayoritas kelas dari k data terdekat di sekitarnya. KNN kerap dimanfaatkan dalam analisis data

lingkungan, termasuk dalam pengelompokan data terkait kualitas air. (Ulum et al., 2023). Euclidean distance adalah ukuran jarak antara dua titik dalam suatu ruang, yang dihitung menggunakan rumus Pythagoras. Jarak ini merepresentasikan panjang garis lurus yang menghubungkan dua titik, yaitu titik a dan titik b. Garis tersebut—sering disebut sebagai garis miring—membentang melintasi sumbu x dan sumbu y, berdasarkan koordinat yang dimiliki oleh masing-masing titik. (Malik Namus Akbar, 2024). Berikut menunjukan rumus perhitungan untuk mencari jarak antara dua titik yaitu titik pada data training (x) dan titik pada data testing (y) maka digunakan rumus Euclidean, seperti pada persamaan berikut:

$$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2 + (x_5 - y_5)^2} + \sqrt{(x_6 - y_6)^2 + (x_7 - y_7)^2 + (x_8 - y_8)^2} \quad (1)$$

Dengan D adalah jarak antara titik pada data training x dan titik data testing yang akan diklasifikasi, dimana x = x1,x2,...,xi dan y1,y2,...yi dan I mempresentasikan nilai atribut serta n merupakan dimensi atribut.

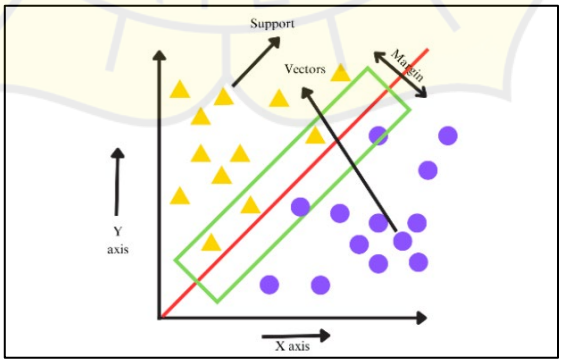
Berikut adalah tahapan dalam menghitung algoritma K-Nearest Neighbor (KNN):

1. Menentukan nilai K, yaitu jumlah tetangga terdekat yang akan digunakan dalam proses klasifikasi.
2. Menghitung jarak Euclidean kuadrat antara data yang ingin diklasifikasikan (query instance) dengan setiap data dalam dataset.
3. Mengurutkan data-data tersebut berdasarkan jarak terdekat, dimulai dari yang memiliki nilai jarak Euclidean terkecil.
4. Mengidentifikasi kategori atau kelas (Y) dari K tetangga terdekat tersebut.

5. Menentukan hasil klasifikasi berdasarkan kategori yang paling banyak muncul di antara tetangga terdekat, dan menggunakan kategori mayoritas tersebut untuk memprediksi kelas dari query instance.

2.1.6 Algoritma *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan algoritma yang digunakan untuk mengelompokkan data, baik yang berpola linear maupun non-linear. Algoritma ini efektif dalam menangani pola non-linear karena menggunakan konsep kernel, yang memungkinkan pemetaan data ke dalam ruang berdimensi lebih tinggi. Dengan SVM, berbagai jenis objek dapat dipisahkan ke dalam kelompok yang berbeda. Sebagai metode supervised learning, SVM melakukan klasifikasi dengan menggunakan data latih yang telah diberi label sebelumnya, untuk memprediksi kelas dari data baru. Kernel linear, yang merupakan jenis kernel paling sederhana, cocok digunakan untuk mengklasifikasikan data yang tidak dapat dipisahkan secara linear dalam ruang aslinya. (Kumala Sari & Randy Suryono, 2024). Ilustrasi SVM digambarkan oleh Gambar 2.2 di bawah ini, yaitu:



Gambar 2. 2 Ilustrasi SVM

Pada Gambar 2.2 ilustrasi SVM di atas, Menjelaskan cara kerja algoritma Support Vector Machine (SVM) dalam memisahkan data ke dalam dua kategori, yaitu

'layak' dan 'tidak layak' untuk dikonsumsi, melalui sebuah garis pemisah yang dikenal sebagai hyperplane. Hyperplane ini ditempatkan sedemikian rupa agar memiliki jarak (margin) maksimum terhadap titik-titik data terdekat dari masing-masing kelas, yang disebut support vectors. Pendekatan ini menghasilkan pemisahan data yang optimal dan akurat. Ilustrasi tersebut merepresentasikan prinsip dasar SVM dalam melakukan klasifikasi kualitas air berdasarkan parameter seperti pH, TDS, kandungan Besi (Fe), Nitrat sebagai N, E. Coli, Total Coliform, warna, dan bau.

Rumus SVM dijelaskan pada persamaan berikut ini, :

$$f(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + w_4 \cdot x_4 + w_5 \cdot x_5 + w_6 \cdot x_6 + w_7 \cdot x_7 + w_8 \cdot x_8 + b \quad (2)$$

Berdasarkan rumus di atas, terdapat penjelasan rumusnya yaitu:

Tabel 2. 1 Penjelasan Rumus Klasifikasi SVM

x_1 = pH	x_6 = Total Coliform
x_2 = TDS (ppm)	x_4 = Warna (hasil encoding numerik)
x_3 = Besi Fe	x_5 = Bau (hasil encoding numerik)
x_4 = Nitrat sebagai N	$w_1 - w_5$ = bobot masing-masing fitur
x_5 = E. Coli	b = bias

2.1.7 Confusion Matrix

Confusion Matrix adalah alat untuk mengukur performa dalam permasalahan klasifikasi pada machine learning, di mana hasil klasifikasinya dapat terdiri dari dua

kelas atau lebih. Matriks ini memuat empat kombinasi utama yang berasal dari perbandingan antara nilai prediksi dan nilai aktual, sehingga dapat memberikan gambaran mengenai keakuratan model dalam melakukan klasifikasi. Di bawah ini, merupakan gambar *Confusion Matrix* :

Tabel 2. 2 *Confusion Matrix*

Class	**Classified as Positive**	**Classified as Negative**
Positif	True Positive (TP)	False Negative (FN)
Negatif	Flase Positive (FP)	True Negative (TN)

Berdasarkan tabel tersebut, dapat dijelaskan bahwa confusion matrix terdiri dari empat komponen utama. True Positive (TP) menunjukkan jumlah data yang diprediksi dengan benar sebagai kategori positif, misalnya air yang layak minum dan memang benar layak. True Negative (TN) menggambarkan jumlah data yang diprediksi dengan benar sebagai kategori negatif, yaitu air yang tidak layak konsumsi dan memang tidak layak. False Positive (FP) terjadi ketika model salah memprediksi data sebagai positif padahal sebenarnya negatif, contohnya air yang tidak layak minum namun diperkirakan layak. Sedangkan False Negative (FN) adalah kondisi di mana data yang seharusnya termasuk kategori positif diprediksi sebagai negatif, seperti air yang sebenarnya layak minum namun diklasifikasikan sebagai tidak layak. (Ulum et al., 2023)

Untuk menentukan model yang paling optimal, penelitian ini perlu menganalisis hasil dari classification report. Di dalam laporan tersebut terdapat metrik evaluasi seperti akurasi, presisi, recall, dan F1-score yang dapat dijadikan dasar dalam

menilai dan memilih model terbaik. Dalam menentukan classification report terdapat persamaan yang dapat dilihat pada tabel 2.3 sebagai berikut:

Tabel 2. 3 Classification Report

Measurement	Definition	Formula
Accuracy (A)	Akurasi digunakan untuk mengetahui seberapa baik sistem dapat mengklasifikasi data dengan benar	$A = \frac{TP + TN}{(Total\ numbers\ of\ sampels)}$
Precision (P)	Seberapa akurat suatu model dapat mengidentifikasi suatu sentimen	$P = \frac{TP}{TP + FP}$
Recall (R)	Seberapa baik model menemukan dan mengenali suatu sentimen	$R = \frac{TP}{TP + FN}$
F1- Score (F)	Gabungan dari Precision dan Recall dimana untuk mengukur performa sebauh sistem	$F = 2\ x\ \frac{P\ x\ R}{P + R}$

Pada tabel 2.3 dalam sebuah classification report, terdapat beberapa metrik penting yang perlu diperhatikan, yaitu accuracy, precision, recall, dan F1-Score. Masing-masing metrik memiliki peran yang berbeda. Accuracy digunakan untuk mengukur sejauh mana algoritma mampu mengklasifikasikan data dengan benar secara keseluruhan. Namun, nilai akurasi yang tinggi saja tidak selalu menunjukkan bahwa suatu algoritma bekerja dengan baik. Oleh karena itu, diperlukan pula analisis terhadap precision, recall, dan F1-Score. Pemilihan metrik evaluasi ini harus disesuaikan dengan karakteristik data yang digunakan, sehingga kombinasi dari berbagai metrik tersebut dapat membantu dalam pengambilan keputusan yang lebih tepat untuk menentukan model algoritma yang paling sesuai.. (Ramadhani & Suryono, 2024)

2.1.8 Pemodelan Sistem UML

Unified Modeling Language (UML) adalah standar pemodelan visual yang digunakan untuk menggambarkan, merancang, serta mendokumentasikan sistem perangkat lunak. Dalam tahap ini, UML dimanfaatkan untuk memodelkan sistem

yang akan dibangun agar lebih mudah dipahami oleh pengembang maupun pemangku kepentingan. Pemodelan tersebut mencakup berbagai jenis diagram, seperti diagram use case, diagram aktivitas, dan diagram kelas, yang masing-masing merepresentasikan aspek tertentu dari sistem. Penggunaan UML membantu menjadikan proses pengembangan sistem lebih sistematis, terorganisir, dan terarah.

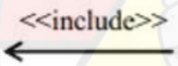
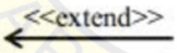
2.1.8.1 Unified Modelling Language (UML)

Unified Modeling Language (UML) merupakan bahasa spesifikasi standar yang digunakan untuk mendokumentasikan, merancang, dan membangun perangkat lunak. UML berperan sebagai metodologi dalam pengembangan sistem berbasis objek, sekaligus menjadi alat bantu yang mendukung proses perancangan dan pengembangan sistem secara efektif dan terstruktur. (Suendri, 2018).

2.1.8.1.1 Use Case Diagram

*Use case diagram a*dalah jenis pemodelan yang digunakan untuk menggambarkan perilaku (behavior) dari sistem informasi yang akan dikembangkan. Diagram ini berfungsi untuk mengidentifikasi berbagai fungsi atau fitur yang terdapat dalam sistem, serta menentukan siapa saja (aktor) yang memiliki hak akses atau berinteraksi dengan fungsi-fungsi tersebut. (Irfan et al., 2023) Simbol-simbol yang digunakan dalam *Use Case* Diagram yaitu:

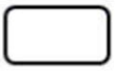
Tabel 2. 4 Simbol *Use case* Diagram (Agung Noviantoroa, 2022)

Simbol	Keterangan
	Aktor : Mewakili peran orang, sistem yang lain, atau alat ketika berkomunikasi dengan *use case*
	*Use case* : Abstraksi dan interaksi antara sistem dan aktor
	*Association* : Abstraksi dari penghubung antara aktor dengan use case
	Generalisasi : Menunjukkan spesialisasi aktor untuk dapat berpartisipasi dengan use case
	Menunjukkan bahwa suatu use case seluruhnya merupakan fungsionalitas dari use case lainnya
	Menunjukkan bahwa suatu use case merupakan tambahan fungsional dari use case lainnya jika suatu kondisi terpenuhi

2.1.8.1.2 *Activity* Diagram

Activity Diagram digunakan untuk memvisualisasikan alur kerja (workflow) atau rangkaian aktivitas dalam suatu sistem maupun proses bisnis. Diagram ini menunjukkan bagaimana proses berlangsung dari satu aktivitas ke aktivitas lainnya secara terstruktur. (Irfan et al., 2023) Simbol-simbol yang digunakan dalam *Activity Diagram* yaitu:

Tabel 2. 5 Simbol *Activity* Diagram (Agung Noviantoroa, 2022)

Simbol	Nama	Keterangan
	Status awal	Sebuah diagram aktivitas memiliki sebuah status awal.
	Aktivitas	Aktivitas yang dilakukan sistem,aktivitas biasanya diawali dengan kata kerja.
	Percabangan / Decision	Percabangan dimana ada pilihan aktivitas yang lebih dari satu.
	Penggabungan / Join	Penggabungan dimana yang mana lebih dari satu aktivitas lalu digabungkan jadi satu.
	Status Akhir	Status akhir yang dilakukan sistem, sebuah diagram aktivitas memiliki sebuah status akhir
	Swimlane	Swimlane memisahkan organisasi bisnis yang bertanggung jawab terhadap aktivitas yang terjadi

2.1.9 Software dan Pemograman Terkait

Perangkat lunak (Software) dan bahasa pemrograman yang digunakan memiliki dampak terhadap efisiensi, kinerja, dan implementasi sistem yang dikembangkan. Dalam hal ini, perangkat lunak yang digunakan mencakup editor untuk pengolahan data serta bahasa pemrograman yang relevan dengan kebutuhan penelitian.

2.1.9.1 Editor Google Colabs

Google Colab adalah platform yang memungkinkan pengguna untuk menulis dan menjalankan kode Python langsung dari browser, sehingga sangat cocok untuk keperluan machine learning dan analisis data. Layanan ini menyediakan akses ke sumber daya komputasi yang kuat dan mempermudah proses penyimpanan serta berbagi notebook secara online. (Fadil Danu Rahman et al., 2024).

2.1.9.2 Pemograman Python

Python merupakan bahasa pemrograman serbaguna yang memiliki banyak keunggulan, sehingga menjadi pilihan utama di berbagai bidang seperti pengembangan web, data science, hingga kecerdasan buatan. Meskipun Python memiliki beberapa kekurangan, seperti performa yang cenderung lebih lambat dan keterbatasan dalam multithreading, manfaat yang ditawarkannya sering kali jauh lebih besar dibandingkan kelemahannya. Python didukung oleh komunitas yang luas, koleksi library yang lengkap, serta sintaksis yang sederhana dan mudah dipahami, menjadikannya tetap relevan dan kuat dalam pengembangan teknologi modern. Beberapa library populer yang mendukung kemampuan Python antara lain NumPy untuk komputasi ilmiah, Pandas untuk analisis data, Matplotlib dan Seaborn untuk visualisasi, Scikit-Learn untuk machine learning, serta TensorFlow, Keras, dan PyTorch untuk deep learning dan kecerdasan buatan. Dalam pengembangan web, Python juga diperkuat oleh framework seperti Django dan Flask yang memungkinkan pembuatan aplikasi web secara cepat dan efisien..

(Junaidi et al., 2023). Pada pemrograman python, terdapat beberapa library yang digunakan yaitu:

2.1.9.2.1 Library Pandas

Pandas adalah pustaka open-source yang sangat populer dalam bahasa pemrograman Python, dirancang khusus untuk analisis dan manipulasi data. Pandas menyediakan berbagai alat yang efisien dan fleksibel untuk menangani data terstruktur, sehingga memudahkan pengguna dalam proses pengolahan dan analisis data secara menyeluruh. (Prof. Dr. Moechammad Sarosa, 2022). Untuk mengimport paket atau library panda dapat dilakukan dengan perintah berikut:

```
import pandas as pd

df = pd.read_csv('data.csv')

print(df.to_string())
```

Gambar 2. 3 Library Panda (sumber: w3schools)

Dalam proses pembersihan dan prapemrosesan data, Pandas menyediakan beragam fungsi yang berguna untuk menangani data yang hilang, menghapus data duplikat, serta melakukan transformasi data. Fitur-fitur ini membantu peneliti dalam menyiapkan data secara tepat dan memastikan kualitas data yang baik sebelum dianalisis lebih lanjut. (Lavanya et al., 2023).

2.1.9.2.2 Library NumPy

NumPy (Numerical Python) merupakan pustaka dalam bahasa Python yang dirancang khusus untuk melakukan komputasi numerik. Library ini

menyediakan objek array berdimensi-N (N-dimensional array) yang sangat efisien untuk mendukung berbagai operasi matematika dan analisis data. (Prof. Dr. Moechammad Sarosa, 2022). Pertama yang harus dilakukan untuk memulai suatu project menggunakan library Numpy yaitu sebagai berikut:

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5])

print(arr)

print(type(arr))
```

Gambar 2. 4 Library NumPy (sumber: w3schools)

NumPy memiliki sejumlah fitur utama, seperti menyediakan fungsi-fungsi numerik yang cepat dan sudah terkompilasi, mendukung komputasi berbasis array secara efisien, serta mendukung pendekatan berorientasi objek. Selain itu, NumPy memungkinkan proses vektorisasi, yang membuat perhitungan menjadi lebih singkat dan cepat. NumPy juga sangat berguna dalam analisis data, karena memungkinkan pembuatan array N-dimensi yang kuat, serta menjadi dasar bagi pustaka lain seperti SciPy dan scikit-learn. (Salah & Din, 2020).

2.1.9.2.3 Library Matplotlib

Matplotlib merupakan salah satu pustaka visualisasi data yang paling dikenal dalam bahasa pemrograman Python. Dikembangkan oleh John Hunter dan para kontributor lainnya, Matplotlib dirancang khusus untuk mendukung kebutuhan

para ilmuwan dan peneliti di seluruh dunia. Sebagai bagian penting dalam ekosistem data science Python, Matplotlib terintegrasi dengan baik bersama pustaka lain seperti NumPy dan Pandas, sehingga menjadi alat esensial dalam proses analisis data. Library ini memungkinkan pengguna membuat grafik interaktif yang dapat dimodifikasi secara langsung, mendukung eksplorasi data serta mempermudah dalam menemukan pola dan tren. Matplotlib juga mendukung pembuatan visualisasi dengan kualitas tinggi yang cocok untuk publikasi ilmiah, presentasi, maupun materi akademik lainnya. Kemampuannya untuk bekerja selaras dengan pustaka seperti NumPy, Pandas, dan SciPy menjadikan Matplotlib sangat berguna bagi peneliti dalam menganalisis dan menyampaikan hasil temuannya melalui grafik yang jelas dan informatif. (Hafeez & Sial, 2021).

```
import matplotlib.pyplot as plt
import numpy as np

xpoints = np.array([0, 6])
ypoints = np.array([0, 250])

plt.plot(xpoints, ypoints)
plt.show()
```

Gambar 2. 5 Library Matplotib (sumber: w3schools)

2.1.9.2.4 Library Scikit-learn

Scikit-learn merupakan pustaka open source yang sangat populer dalam analisis data dan menjadi acuan utama untuk penerapan Machine Learning (ML) di ekosistem Python. Dalam penelitian ini, proses pembangunan model, pelatihan (training), serta pengujian (testing) dilakukan untuk memperoleh nilai akurasi dan

prediksi yang optimal dengan menerapkan algoritma KNN dan SVM. Seluruh tahapan pengolahan data dilakukan menggunakan bahasa pemrograman Python melalui platform Google Colab. (Saputra et al., 2023). Pustaka ini mencakup berbagai metode algoritma dalam data mining, seperti algoritma untuk klasifikasi, regresi, dan pengelompokan (clustering). (ActiveState, 2021).

```
from sklearn import datasets

X, y = datasets.load_iris(return_X_y=True)
```

Gambar 2. 6 Library Scikit-learn (sumber: w3schools)

2.1.9.2.5 Library Seaborn

Seaborn merupakan pustaka Python yang digunakan untuk membuat grafik statistik, dibangun sebagai antarmuka tingkat tinggi di atas Matplotlib dan terintegrasi erat dengan struktur data dari Pandas. Library ini menyederhanakan proses pembuatan visualisasi informatif dari data, melalui API deklaratif yang berfokus pada dataset. Saat pengguna memberikan data dan jenis plot yang diinginkan, Seaborn secara otomatis memetakan nilai-nilai tersebut ke elemen visual seperti warna, ukuran, dan gaya, serta melakukan transformasi statistik yang dibutuhkan. Selain itu, Seaborn melengkapi visualisasi dengan label sumbu dan legenda yang jelas. Banyak fungsinya memungkinkan pembuatan grafik multi-panel, yang berguna untuk membandingkan subset data atau pasangan variabel yang berbeda. Seaborn sangat bermanfaat dalam setiap tahap proyek analisis ilmiah—mulai dari eksplorasi awal hingga pembuatan grafik akhir—karena mampu menghasilkan visualisasi kompleks dengan sintaks yang ringkas. Ditambah dengan

berbagai pilihan kustomisasi dan akses ke objek Matplotlib, Seaborn juga memungkinkan pembuatan visualisasi yang profesional dan siap untuk publikasi. (Waskom, 2021).

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.displot([0, 1, 2, 3, 4, 5])

plt.show()
```

Gambar 2. 7 Library Seaborn (sumber: w3schools)

2.1.9.2.6 Streamlit

Streamlit adalah framework open-source yang dirancang untuk memudahkan pembuatan aplikasi web menggunakan bahasa pemrograman Python. Dengan Streamlit, pengembang dapat membangun antarmuka web secara cepat hanya dengan skrip Python sederhana, tanpa perlu menguasai pengembangan frontend secara mendalam. Framework ini memungkinkan pembuatan UI yang interaktif dan responsif, lengkap dengan visualisasi data, tabel, grafik, serta elemen interaktif lainnya. Selain itu, Streamlit juga menyediakan kemudahan dalam proses deployment, sehingga aplikasi yang dibuat dapat diakses secara online dengan mudah oleh pengguna lainnya.

2.1.10 Waterfall

Pada perancangan sistem ini, digunakan pendekatan pengembangan dengan metode System Development Life Cycle (SDLC), khususnya model Waterfall. Waterfall

merupakan metode pengembangan sistem yang prosesnya dilakukan secara bertahap dan berurutan dari satu fase ke fase berikutnya. Menurut Sommerville (2011: p30), model Waterfall terdiri dari beberapa tahapan, yaitu:

a. Requirements analysis and definition

Merupakan tahap penentuan fitur, batasan, serta tujuan dari sistem yang dilakukan melalui diskusi dengan para pengguna. Seluruh aspek tersebut dirumuskan secara mendetail dan akan menjadi acuan spesifikasi sistem yang akan dikembangkan..

b. System and software design

Pada tahap ini, arsitektur sistem dirancang berdasarkan kebutuhan yang telah ditentukan sebelumnya. Selain itu, dilakukan pula identifikasi dan pemodelan terhadap abstraksi dasar dari sistem perangkat lunak beserta relasi antar komponennya.

c. Implementation and unit testing

Pada tahap ini, rancangan perangkat lunak yang telah dibuat akan diimplementasikan menjadi kumpulan program atau unit-unit program. Masing-masing unit tersebut kemudian akan diuji untuk memastikan bahwa semuanya sesuai dengan spesifikasi yang telah ditetapkan.

d. Integration and system testing

Pada tahap ini, seluruh unit program digabungkan dan diuji secara keseluruhan sebagai satu kesatuan sistem untuk memastikan bahwa sistem telah sesuai dengan

persyaratan yang ditentukan. Setelah proses pengujian selesai, sistem kemudian diserahkan kepada pengguna.

e. Operation and maintenance

Pada tahap ini, sistem dipasang dan mulai dioperasikan oleh pengguna. Selain itu, kesalahan yang belum terdeteksi saat proses pengembangan juga diperbaiki. Tahap ini juga mencakup pengembangan lanjutan, seperti penambahan fitur serta peningkatan fungsi sistem.

2.2 Tinjauan Literatur/Kajian Penelitian Terdahulu

Tinjauan literatur atau kajian penelitian sebelumnya dilakukan untuk membangun dasar teori yang kuat dan memastikan bahwa penelitian ini memberikan kontribusi baru dalam bidang klasifikasi kualitas air. Kajian ini mencakup berbagai penelitian terkait penerapan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) dalam klasifikasi data, serta pemanfaatan data mining untuk mendukung proses penilaian dan prediksi kelayakan kualitas air sumur, khususnya di wilayah Jakarta Timur.

2.2.1 Paper 1

Judul: Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma *K-Nearest Neighbor* dan Logistic Regression

Author: Baiq Nurul Azmi1, Arief Hermawan, Donny Avianto

Publikasi: Jurnal Sistem dan Teknologi Informasi (JUSTINDO)

Tahun: 2022

Klasifikasi Journal: Sinta 3

2.2.1.1 Tujuan Penelitian

Penelitian ini bertujuan untuk menganalisis kualitas air di lingkungan perkotaan dengan menggunakan metode *K-Nearest Neighbor* (kNN) dan Regresi Logistik, serta mengevaluasi pengaruh Principal Component Analysis (PCA) terhadap akurasi klasifikasi. Penelitian ini akan mengidentifikasi fitur-fitur yang paling berpengaruh dalam klasifikasi kualitas air dan menentukan performa akurasi dari kedua metode klasifikasi yang digunakan. Selain itu, penelitian ini juga akan membandingkan hasil klasifikasi sebelum dan sesudah penerapan PCA untuk memahami dampaknya terhadap akurasi. Dengan demikian, penelitian ini diharapkan dapat memberikan rekomendasi untuk penelitian lebih lanjut dalam pengembangan algoritma klasifikasi yang lebih efisien dan efektif dalam penilaian kualitas air.

2.2.1.2 Metodelogi Yang Digunakan

Penelitian ini menggunakan algoritma *K-Nearest Neighbor* (KNN) dan Regresi Logistik untuk menganalisis kualitas air di lingkungan perkotaan. Metodologi yang diterapkan mencakup pengumpulan data sekunder dari dataset kualitas air yang tersedia di Kaggle, yang terdiri dari 8.000 sampel dengan 21 atribut. Selanjutnya, dilakukan pra-pemrosesan data untuk memastikan kualitas dan konsistensi data. Penelitian ini kemudian mengevaluasi performa kedua metode klasifikasi dengan membagi data menjadi data latih dan data uji dengan proporsi 80%:20%. Setelah itu, dilakukan pengujian untuk membandingkan hasil klasifikasi sebelum dan sesudah penerapan Principal Component Analysis (PCA) sebagai

langkah reduksi dimensi. Hasil analisis ini diharapkan dapat memberikan wawasan yang lebih baik mengenai akurasi klasifikasi kualitas air dan rekomendasi untuk pengembangan algoritma yang lebih efisien di masa mendatang.

2.2.1.3 Temuan Utama

Penelitian ini menemukan bahwa penerapan metode *K-Nearest Neighbor* (kNN) dan Regresi Logistik berhasil mengklasifikasikan kualitas air dengan akurasi tertinggi mencapai 90.8% untuk kNN tanpa penerapan Principal Component Analysis (PCA). Hasil analisis menunjukkan bahwa penggunaan PCA cenderung menurunkan performa akurasi kedua metode, di mana akurasi kNN dengan PCA mencapai 88.5%. Temuan ini menegaskan pentingnya pemilihan fitur yang tepat dalam klasifikasi kualitas air dan memberikan wawasan bagi pengembangan algoritma klasifikasi yang lebih efisien di masa mendatang, serta menyoroti urgensi untuk terus memantau dan mengelola kualitas air di lingkungan perkotaan.

2.2.1.4 Kesimpulan Penelitian

Metode *K-Nearest Neighbor* (kNN) terbukti lebih efektif dalam mengklasifikasikan kualitas air dibandingkan dengan Logistic Regression, dengan akurasi tertinggi mencapai 90.8% tanpa penerapan Principal Component Analysis (PCA). Penerapan PCA justru menurunkan performa kedua metode, di mana akurasi kNN menjadi 88.5% dan Logistic Regression menjadi 87.9%. Temuan ini menunjukkan bahwa penggunaan banyak fitur dalam klasifikasi kualitas air sangat penting untuk mencapai akurasi yang tinggi. Penelitian ini memberikan wawasan yang berharga untuk pengembangan algoritma klasifikasi yang lebih efisien dan

efektif dalam penilaian kualitas air, serta menekankan perlunya pemantauan yang berkelanjutan terhadap kualitas air di lingkungan perkotaan. (Azmi et al., 2022)

2.2.2 Paper 2

Judul: Metode KNN (*K-Nearest Neighbor*) untuk Menentukan Kualitas Air

Author: Fahmi Malik Namus Akbar

Publikasi: Jurnal TEKNO KOMPAK

Tahun: 2024

Klasifikasi Journal: Sinta 4

2.2.2.1 Tujuan Penelitian

Penelitian ini bertujuan untuk menerapkan algoritma *K-Nearest Neighbor* (KNN) dalam mengklasifikasikan kualitas air berdasarkan 20 unsur kimia yang relevan. Penelitian ini diharapkan dapat memberikan wawasan baru mengenai efektivitas metode KNN dalam menentukan kualitas air, serta membantu masyarakat dan pemerintah dalam mengambil tindakan respons yang lebih efisien terkait potensi kontaminasi dan masalah kualitas air lainnya.

2.2.2.2 Metodelogi Yang Digunakan

Metodologi penelitian ini meliputi pengumpulan data sekunder dari dataset yang diperoleh dari situs Kaggle.com, yang berjudul "waterQuality1", yang berisi informasi mengenai 20 unsur kimia yang relevan untuk menentukan kualitas air. Setelah pengumpulan data, tahap pra-pemrosesan dilakukan untuk memastikan bahwa data siap digunakan, termasuk penanganan nilai yang hilang dan normalisasi

data. Selanjutnya, algoritma *K-Nearest Neighbor* (KNN) diterapkan untuk mengklasifikasikan sampel air berdasarkan fitur-fitur yang ada, dengan menggunakan data latih yang terdiri dari 800 record. Proses klasifikasi melibatkan pencarian tetangga terdekat dan penentuan label mayoritas dari tetangga tersebut. Setelah model klasifikasi terbentuk, dilakukan evaluasi akurasi menggunakan data uji untuk menilai efektivitas metode KNN. Hasil analisis disusun dalam laporan untuk memberikan wawasan mengenai kualitas air dan potensi kontaminasi, serta untuk mendukung tindakan respons yang lebih efisien terkait masalah kualitas air.

2.2.2.3 Temuan Utama

Penelitian ini menemukan bahwa metode *K-Nearest Neighbor* (KNN) efektif dalam mengklasifikasikan kualitas air berdasarkan 20 unsur kimia yang relevan. Hasil klasifikasi menunjukkan bahwa sampel air dapat dibagi menjadi kategori aman dan tidak aman untuk dikonsumsi, dengan tingkat akurasi mencapai 92% pada data uji. Penelitian ini juga mengidentifikasi unsur-unsur kimia tertentu, seperti Arsenic dan Bakteri, yang memiliki dampak signifikan terhadap kualitas air. Selain itu, penggunaan dataset dari Kaggle memberikan keandalan dalam analisis, dan hasil penelitian ini diharapkan dapat membantu masyarakat dan pemerintah dalam mengambil tindakan respons yang lebih efisien terkait potensi kontaminasi air.

2.2.2.4 Kesimpulan Penelitian

Penerapan metode *K-Nearest Neighbor* (KNN) dalam klasifikasi kualitas air menggunakan dataset dari Kaggle menunjukkan hasil yang signifikan. Penelitian ini berhasil mengklasifikasikan sampel air menjadi kategori aman dan tidak aman

untuk dikonsumsi dengan tingkat akurasi mencapai 92%. Hasil analisis mengindikasikan bahwa unsur-unsur kimia tertentu, seperti Arsenic dan Bakteri, memiliki pengaruh besar terhadap kualitas air. Selain itu, penelitian ini menekankan pentingnya pemantauan kualitas air untuk kesehatan masyarakat dan pengelolaan sumber daya air yang lebih baik. Dengan demikian, KNN terbukti sebagai metode yang efektif dalam menentukan kualitas air, memberikan wawasan berharga bagi masyarakat dan pemerintah dalam menangani isu-isu terkait kualitas air. (Malik Namus Akbar, 2024).

2.2.2 Paper 3

Judul: Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest

Author: Muhamad Malik Mutoffar, Muchammad Naseer, Ariansyah Fadillah

Publikasi: Jurnal Nasional Riset Aplikasi dan Teknik Informatika

Tahun: 2022

Klasifikasi Journal: Sinta 4

2.2.3.1 Tujuan Penelitian

Penelitian ini bertujuan untuk menerapkan algoritma Random Forest dalam mengklasifikasikan kualitas air tanah di Jakarta dan untuk menentukan kondisi kualitas air tanah yang layak atau tidak layak konsumsi.

2.2.3.2 Metodelogi Yang Digunakan

Metode yang digunakan dalam penelitian ini adalah algoritma Random Forest untuk klasifikasi kualitas air sumur. Tahapan penelitian mencakup pengumpulan data, pengolahan data, implementasi algoritma, dan evaluasi hasil.

Data yang digunakan terdiri dari 267 sampel yang dianalisis melalui software Orange dengan pembagian 80% untuk pelatihan dan 20% untuk pengujian.

2.2.3.3 Temuan Utama

Temuan utama dari penelitian ini menunjukkan bahwa algoritma Random Forest mampu memberikan akurasi yang baik dalam klasifikasi kualitas air. Hasil pengujian menghasilkan presisi sebesar 0,823 dan sensitivitas sebesar 0,83, yang menunjukkan tingkat keakuratan yang cukup tinggi dalam memprediksi air yang layak konsumsi dan yang tidak layak konsumsi.

2.2.3.4 Kesimpulan Penelitian

Kesimpulan dari penelitian ini adalah bahwa algoritma Random Forest efektif dalam mengklasifikasikan kualitas air tanah di Jakarta. Dengan presisi 82,3% dan sensitivitas 83%, model ini mampu memberikan prediksi yang cukup baik, membantu dalam menentukan kelayakan air tanah untuk konsumsi. (Mutoffar & Fadillah, 2022).

2.2.4 Paper 4

Judul: Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree

Author: Aldi Tangkelayuk, Evangs Mailoa

Publikasi: Jurnal Nasional Riset Aplikasi dan Teknik Informatika

Tahun: Jurnal Teknik Informatika dan Sistem Informasi

Klasifikasi Journal: Sinta 4

2.2.4.1 Tujuan Penelitian

Penelitian ini bertujuan untuk membandingkan tingkat akurasi dari tiga metode klasifikasi—*K-Nearest Neighbor*s (KNN), Naïve Bayes, dan Decision Tree—dalam mengklasifikasikan kualitas air, guna menentukan metode yang paling akurat untuk identifikasi kualitas air layak konsumsi.

2.2.4.2 Metedologi Yang Digunakan

Metode yang digunakan adalah *K-Nearest Neighbor*s, Naïve Bayes, dan Decision Tree. Penelitian dilakukan dengan dataset kualitas air dari Kaggle, yang diolah menggunakan software Rapid Miner. Dataset dibagi menjadi data pelatihan (70%) dan data pengujian (30%). Pengukuran akurasi dilakukan menggunakan *Confusion Matrix* untuk setiap metode, yang memberikan nilai akurasi sebagai perbandingan performa algoritma.

2.2.4.3 Temuan Utama

Temuan utama penelitian ini menunjukkan bahwa metode *K-Nearest Neighbor*s memiliki tingkat akurasi tertinggi sebesar 86,88%, dibandingkan dengan Decision Tree sebesar 80,84% dan Naïve Bayes sebesar 63,60%. Hal ini menempatkan *K-Nearest Neighbor*s sebagai metode paling efektif untuk klasifikasi kualitas air di antara ketiga metode yang diuji.

2.2.4.4 Kesimpulan Penelitian

Kesimpulan dari penelitian ini adalah bahwa metode *K-Nearest Neighbor*s merupakan yang paling baik dalam klasifikasi kualitas air karena menghasilkan akurasi tertinggi. Namun, Decision Tree juga memiliki performa yang cukup baik

dengan akurasi di atas 80%, sedangkan Naïve Bayes memiliki akurasi yang paling rendah. Untuk penelitian selanjutnya, disarankan penggunaan lebih banyak data dan metode lain guna meningkatkan akurasi klasifikasi. (Tangkelayuk, 2022)

2.2.5 Paper 5

Judul: Penerapan Algoritma Data Mining untuk Klasifikasi Kualitas Air

Author: Fauzi Yusa Rahman, Indu Indah Purnomo, Nadya Hijriana

Publikasi: Technologia, Jurnal Ilmiah

Tahun: 2022

Klasifikasi Journal: Sinta 4

2.2.5.1 Tujuan Penelitian

Penelitian ini bertujuan untuk menerapkan metode data mining dalam klasifikasi kualitas air dengan menguji performa algoritma Decision Tree, Naive Bayes, dan *K-Nearest Neighbor*, guna mengetahui algoritma yang paling akurat untuk mengklasifikasi kualitas air berdasarkan parameter mikrobiologi, kimia anorganik, dan kimia.

2.2.5.2 Metedologi Yang Digunakan

Penelitian ini menggunakan metode eksperimen dengan data kualitas air dari Kaggle. Proses pengujian dilakukan dengan membagi data menjadi data latih dan data uji menggunakan teknik K-fold cross-validation sebanyak 10 kali. Kinerja setiap algoritma dievaluasi menggunakan *Confusion Matrix* dan kurva ROC untuk menilai akurasi dan kemampuan klasifikasi dari setiap model.

2.2.5.3 Temuan Utama

Temuan utama penelitian ini menunjukkan bahwa algoritma Decision Tree memiliki akurasi tertinggi sebesar 94,94% dan nilai AUC 0,865, yang menjadikannya algoritma terbaik dalam klasifikasi kualitas air. Metode *K-Nearest Neighbor* mengikuti dengan akurasi 87,86%, sedangkan Naive Bayes menghasilkan akurasi sebesar 84,79%.

2.2.5.4 Kesimpulan Penelitian

Kesimpulan dari penelitian ini adalah bahwa algoritma Decision Tree paling unggul dalam mengklasifikasi kualitas air dengan akurasi tertinggi, termasuk dalam golongan klasifikasi yang baik. Algoritma ini efektif untuk penentuan kualitas air yang aman dikonsumsi, sehingga dapat menjadi solusi dalam mengidentifikasi kualitas air berdasarkan parameter yang diteliti. (Rahman et al., 2022)