**PENERAPAN SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI STANDAR KELAYAKAN AIR MINUM DENGAN SKEMA OPTIMALISASI PARAMETER**

**USULAN SKRIPSI**

**Oleh :**

**BAYU PERMANA YOGA**

**NIM. 200605110145**

****

**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM**

**MALANG**

**2025**

# HALAMAN PERSETUJUAN

# DAFTAR ISI

[HALAMAN PERSETUJUAN i](#_Toc219195884)

[DAFTAR ISI ii](#_Toc219195885)

[DAFTAR TABEL iv](#_Toc219195886)

[DAFTAR GAMBAR v](#_Toc219195887)

[BAB I PENDAHULUAN 1](#_Toc219195888)

[1.1 Latar Belakang 1](#_Toc219195889)

[1.2 Rumusan Masalah 3](#_Toc219195890)

[1.3 Tujuan Penelitian 3](#_Toc219195891)

[1.4 Batasan Masalah 4](#_Toc219195892)

[1.5 Manfaat Penelitian 4](#_Toc219195893)

[BAB II STUDI PUSTAKA 5](#_Toc219195894)

[2.1 Penelitian Terkait 5](#_Toc219195895)

[2.2 Kelayakan Air Minum 11](#_Toc219195896)

[2.3 Support Vector Machine 13](#_Toc219195897)

[BAB III METODOLOGI PENELITIAN 15](#_Toc219195898)

[3.1 Metodologi Penelitian 15](#_Toc219195899)

[3.2 Pengumpulan Data 16](#_Toc219195900)

[3.3 Desain Sistem 18](#_Toc219195901)

[3.3.1 Pengumpulan data 18](#_Toc219195902)

[3.3.2 Preprocessing 20](#_Toc219195903)

[3.3.3 Dataset 20](#_Toc219195904)

[3.3.4 Data Split 21](#_Toc219195905)

[3.3.5 Tuning Parameter (GridSearch, inisialisasi C, gamma, kernel, CV) 21](#_Toc219195906)

[3.3.6 Pelatihan 23](#_Toc219195907)

[3.3.7 Testing 23](#_Toc219195908)

[3.3.8 Evaluasi 23](#_Toc219195909)

[3.4 Skenario Pengujian 24](#_Toc219195910)

[3.4.1 Normalisasi Data dengan Z-Score Standardization 24](#_Toc219195911)

[3.4.2 Perhitungan Support Vector Machine Kernel RBF 26](#_Toc219195912)

[3.4.3 Perhitungan Metrik Evaluasi 28](#_Toc219195913)

[BAB IV UJI COBA DAN PEMBAHASAN 30](#_Toc219195914)

[4.1 Lingkungan Implementasi 30](#_Toc219195915)

[4.1.1 Perangkat Keras (Hardware) : 30](#_Toc219195916)

[4.1.2 Perangkat Lunak (Software): 30](#_Toc219195917)

[4.2 Deskripsi dan Persiapan Data 31](#_Toc219195918)

[4.2.1 Analisis Keseimbangan Kelas 33](#_Toc219195919)

[4.2.2 Pra-pemrosesan Data 34](#_Toc219195920)

[4.3 Hasil Skenario Pengujian 35](#_Toc219195921)

[4.3.1 Hasil *Hyperparameter Tuning* (GridSearchCV) 35](#_Toc219195922)

[4.3.2 Perbandingan Kinerja Model pada Data Uji 38](#_Toc219195923)

[4.3.3 Visualisasi Confusion Matrix 38](#_Toc219195924)

[4.4 Pembahasan (Diskusi) 41](#_Toc219195925)

[4.4.1 Analisis Model Terbaik dan Pengaruh Pembagian Data 41](#_Toc219195926)

[4.4.2 Efektivitas Penanganan *Class Imbalance* 41](#_Toc219195927)

[4.4.3 Jawaban atas Rumusan Masalah 42](#_Toc219195928)

[4.5 Implementasi Model Akhir 42](#_Toc219195929)

[BAB V KESIMPULAN 43](#_Toc219195930)

[5.1 Kesimpulan 43](#_Toc219195931)

[5.2 Saran 43](#_Toc219195932)

[DAFTAR PUSTAKA 45](#_Toc219195933)

# DAFTAR TABEL

[Tabel 3.1 Penjelasan Data Kualitas Air 23](#_Toc219196482)

[Tabel 3.2 Dataset Kualitas Air 23](#_Toc219196483)

[Tabel 3.3 Skenario Pengujian 30](#_Toc219196484)

[Tabel 3.4 Statistik Dasar Data 31](#_Toc219196485)

[Tabel 3.5 Perhitungan Manual *Standard Scaling* 31](#_Toc219196486)

[Tabel 3.6 hasil normalisasi menggunakan *standard scaling* 32](#_Toc219196487)

[Tabel 3.7 Perhitungan manual kernel RBF 33](#_Toc219196488)

[Tabel 3.8 perhitungan Kernel RBF 33](#_Toc219196489)

[Tabel 3.9 Matrix evaluasi 34](#_Toc219196490)

[Tabel 4.1 Distribusi Data 37](#_Toc219196491)

[Tabel 4.2 Detail Iterasi Skenario 60:40 41](#_Toc219196492)

[Tabel 4.3 Detail Iterasi 70:30 42](#_Toc219196493)

[Tabel 4.4 Detail Iterasi 60:40 43](#_Toc219196494)

# DAFTAR GAMBAR

[Gambar 2.1 Ilustrasi Support Vector Machine 20](#_Toc219196468)

[Gambar 3.1 Metodologi Penelitian 21](#_Toc219196469)

[Gambar 3.2 Desain Sistem klasifikasi kelayakan air minum dengan SVM 24](#_Toc219196470)

[Gambar 4.1 Distribusi Fitur Parameter Kualitas Air 38](#_Toc219196471)

[Gambar 4.2 Matriks Korelasi Antar Fitur 39](#_Toc219196472)

[Gambar 4.3 Distribusi Kelas Layak dan Tidak Layak 40](#_Toc219196473)

[Gambar 4.4 Confusion Matriks Skenario 70:30 44](#_Toc219196474)

[Gambar 4.5 Kurva ROC Skenario 70:30 45](#_Toc219196475)

# BAB I PENDAHULUAN

## Latar Belakang

Air merupakan kebutuhan dasar bagi kehidupan manusia, berfungsi vital tidak hanya untuk keperluan rumah tangga, tetapi juga menyokong berbagai sektor seperti pertanian, industri, transportasi, dan pariwisata (Atmaja, 2019). Inti dari kebutuhan ini adalah ketersediaan air minum yang layak konsumsi, yang harus memenuhi standar kualitas tertentu agar tidak menimbulkan risiko negatif terhadap kesehatan. Sayangnya, masalah ini masih signifikan di Indonesia; berdasarkan data Badan Pusat Statistik (BPS) tahun 2023, sekitar 18% penduduk Indonesia masih mengonsumsi air yang tidak memenuhi standar kelayakan, baik dari sisi parameter fisik, kimia, maupun mikrobiologi. Kondisi air minum yang tidak layak ini berpotensi besar menyebabkan berbagai penyakit menular seperti diare, kolera, dan gangguan kesehatan serius lainnya.

Untuk mengatasi tantangan tersebut, penentuan kelayakan air minum memerlukan metode evaluasi yang cepat dan efisien. Dalam konteks ini, parameter fisik menjadi indikator awal yang krusial. Parameter seperti pH, salinitas, kekeruhan, suhu, dan *Total Dissolved Solids* (TDS) dapat diukur dengan alat sederhana tanpa memerlukan pengujian laboratorium yang memakan waktu dan biaya besar. Pendekatan berbasis parameter fisik ini sangat penting dan relevan, terutama di wilayah terpencil yang sulit menjangkau fasilitas laboratorium, sehingga memungkinkan deteksi awal kelayakan air minum dapat dilakukan secara cepat dan akurat.

Sejalan dengan tuntutan efisiensi dan akurasi, pemanfaatan teknologi *machine learning* (pembelajaran mesin) menjadi solusi modern dalam klasifikasi kualitas air. Algoritma yang dikenal unggul dalam tugas klasifikasi, seperti *Support Vector Machine* (SVM), telah banyak dimanfaatkan dalam analisis data di berbagai bidang. SVM bekerja dengan mencari *hyperplane* optimal untuk memisahkan data ke dalam kelas-kelas yang berbeda, sehingga sangat sesuai untuk menentukan apakah data air (berdasarkan parameter fisiknya) memenuhi standar kelayakan. Pemanfaatan teknologi ini memungkinkan deteksi kualitas air dilakukan secara cepat dan akurat, kualitas air juga harus memperhatikan kesehatan dan kelayakan konsumsi masyarakat.

Lebih dari sekadar persoalan teknis dan kesehatan, pentingnya akses terhadap air bersih dan layak konsumsi juga tercermin dalam perspektif universal dan agama, khususnya Islam, di mana air memiliki kedudukan yang sangat mulia. Dalam Al-Qur’an disebutkan,

اَوَلَمْ يَرَ الَّذِيْنَ كَفَرُوْٓا اَنَّ السَّمٰوٰتِ وَالْاَرْضَ كَانَتَا رَتْقًا فَفَتَقْنٰهُمَاۗ وَجَعَلْنَا مِنَ الْمَاۤءِ كُلَّ شَيْءٍ حَيٍّۗ اَفَلَا يُؤْمِنُوْنَ ۝٣٠

“*Apakah orang-orang kafir tidak mengetahui bahwa langit dan bumi, keduanya, dahulu menyatu, kemudian Kami memisahkan keduanya dan Kami menjadikan segala sesuatu yang hidup berasal dari air? Maka, tidakkah mereka beriman?*” (QS. Al-Anbiya: 30)

Kemuliaan dan urgensi air dalam kehidupan manusia tidak hanya ditegaskan dari sisi ajaran agama, tetapi juga tercermin dalam upaya nyata untuk menjaga dan memastikan kualitasnya. Ayat ini menegaskan bahwa air adalah sumber utama kehidupan yang harus dijaga dan dikelola dengan baik. Selain itu, Rasulullah SAW juga bersabda, “Manusia mempunyai hak yang sama dalam tiga hal: air, rumput, dan api.” (HR. Abu Dawud). Kedua dalil ini memperkuat bahwa akses terhadap air bersih adalah hak dasar manusia, menjadikan upaya menjamin kualitas air sebagai prioritas bersama, baik dari sisi kesehatan masyarakat, kepentingan sosial, maupun ajaran agama.

Sejalan dengan pentingnya menjaga kualitas air, berbagai inovasi teknologi telah dikembangkan untuk mendukung upaya tersebut, salah satunya dengan pemanfaatan metode *machine learning*. Penelitian Maulina Putri et al. (2023) menunjukkan bahwa SVM dapat memberikan akurasi tinggi dalam klasifikasi kualitas air. Contohnya, sebuah penelitian menggunakan sensor pH, kekeruhan, dan TDS yang dihubungkan dengan mikrokontroler ESP32, kemudian data yang diperoleh diklasifikasikan menggunakan SVM dengan tingkat akurasi di atas 90% bahkan hingga 100% (Maulina Putri et al., 2023). Temuan ini menegaskan bahwa penerapan teknologi modern dapat menjadi solusi efektif dan praktis untuk memastikan air yang dikonsumsi masyarakat memenuhi standar kelayakan, sekaligus mewujudkan hak dasar manusia atas air bersih secara nyata.

Meskipun SVM terbukti efektif, masih terdapat tantangan dalam penerapannya, seperti pemilihan kernel yang tepat, penanganan *outlier* dalam data, serta optimasi parameter model untuk meningkatkan akurasi prediksi. Oleh karena itu, penelitian ini akan mengimplementasikan Support Vector Machine (SVM) untuk menentukan kelayakan air minum berdasarkan standar peraturan umum kualitas air, dengan fokus pada parameter fisik yaitu, suhu, kekeruhan, dan TDS; serta satu parameter kimiawi yaitu pH.

Selain berkontribusi dalam peningkatan teknologi analisis kualitas air, penelitian ini juga mendukung pencapaian *Sustainable Development Goals* (SDGs) Nomor 6, yaitu *Clean Water and Sanitation*, yang bertujuan untuk memastikan ketersediaan dan pengelolaan air bersih bagi seluruh masyarakat. Dengan menerapkan metode SVM, diharapkan dapat tercipta solusi yang lebih efisien, akurat, dan berkelanjutan dalam penentuan kelayakan air minum, sehingga dapat membantu pemerintah dan masyarakat dalam menjaga kualitas air yang dikonsumsi.

## Rumusan Masalah

Seberapa efektif model SVM dapat diterapkan dalam mengklasifikasi status kelayakan air minum (layak/tidak layak) berdasarkan data set kualitas air

## Tujuan Penelitian

Menganalisis efektivitas penerapan model *Support Vector Machine* (SVM) untuk mengklasifikasikan status kelayakan air minum (layak dan tidak layak) berdasarkan parameter kualitas air.

## Batasan Masalah

* Data Penelitian terbatas pada beberapa parameter fisik yaitu *Total Dissolved Solids* (TDS), Kekeruhan, Suhu, Salinitas dan satu parameter kimiawi yaitu pH sehingga jika ada parameter mikrobiologi atau parameter kimia lain yang tidak terdeteksi secara fisik, akan mempengaruhi kelayakan air minum.
* Parameter lain seperti parameter mikrobiologi, parameter kimia selain pH, ataupun parameter radioaktif tidak dapat dikumpulkan karena keterbatasan biaya dan alat.

## Manfaat Penelitian

Penelitian ini diharapkan dapat bermanfaat bagi masyarakat yang berkaitan dengan lokasi diambilnya data demi meningkatkan kualitas hidup dan memenuhi *Sustainable Development Goals* (SDGs) nomor 6 yaitu *Clean Water and Sanitation* atau air bersih dan sanitasi layak.

# BAB II STUDI PUSTAKA

## Penelitian Terkait

Ramon et al mengklasifikasikan status gizi bayi menggunakan SVM dengan kernel RBF berdasarkan data antropometri, mencapai akurasi 87%. Studi tersebut menunjukkan keandalan SVM dalam menangani data kesehatan dengan fitur terbatas. Keterkaitan dengan penelitian ini terletak pada penggunaan kernel RBF dan data nonlinear yang serupa, menjadi dasar pemilihan metode SVM untuk klasifikasi parameter fisik air minum. Namun, penelitian ini memperluas cakupan dengan menambahkan sensor ESP32 untuk pengumpulan data *real-time* dan parameter yang berbeda signifikan, di mana Ramon et al. menggunakan data dengan parameter antropometri, sedangkan penelitian ini menggunakan data kelayakan air (Ramon et al., 2022).

Dzakwan dan Subektiningsih melakukan optimalisasi SVM dengan kernel *polynomial* untuk klasifikasi risiko kesehatan ibu hamil. Penelitian tersebut berhasil meningkatkan akurasi dari 60% menjadi 71% setelah melalui tahapan *preprocessing*. Studi ini menggarisbawahi urgensi *preprocessing* data sebagai fondasi untuk normalisasi dan penanganan *outlier* dalam penelitian klasifikasi air minum. Perbedaannya terletak pada fokus penelitian ini terhadap parameter fisik air, bukan parameter klinis (Dzakwan & Subektiningsih, 2023).

Tabel 2.1 Penelitian Terkait

| Penulis | Input | Metode | Output | Hasil | Batasan |
| --- | --- | --- | --- | --- | --- |
| (Ramon et al., 2022) | 1. Tanggal pengukuran 2. Berat badan 3. Tinggi badan | *Support Vector Machine* (SVM) | * Akurasi * Presisi * Recall * F1-score | Dalam penelitian ini, status gizi bayi dibagi menjadi enam kategori yang berbeda. Penggunaan kernel RBF menghasilkan performa klasifikasi yang sangat baik, dengan akurasi, presisi, dan *recall* masing-masing 87%, serta f1-score 86%. | Data set yang digunakan merupakan data lengkap yang diperoleh dari Posyandu Kecamatan Bangunan Purba pada tahun 2020 sebanyak 1732 data. |
| (Dzakwan & Subektiningsih, 2023) | 1. Usia 2. Tekanan darah sistolik 3. Diastolik BP 4. Kadar glukosa 5. Detak jantung 6. Risiko | *Support Vector Machine* (SVM) | * Akurasi * Presisi * Recall * F1-score | Dilakukan pengelompokan risiko kesehatan ibu hamil menjadi tiga kategori. Hasil klasifikasi risiko kematian ibu hamil dengan metode kernel *polynomial* menunjukkan peningkatan ketepatan dari 60% (sebelum pemrosesan data) menjadi 71% (setelah pemrosesan data). | Data set yang digunakan adalah data yang diperoleh dari UCI data set sebanyak 1014 data. |
| (Vidiastanta et al., 2020) | 1. pH level 2. *Total Dissolved Solids* (TDS) 3. Nitrit (NO2) 4. Nitrat (NO3) 5. Kesadahan 6. Klorida 7. Mangan | *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM) | * Akurasi | Penelitian ini membandingkan KNN dan SVM untuk klasifikasi kualitas air. KNN sedikit lebih unggul dengan akurasi 88,94% berbanding 87,71% pada SVM. Parameter optimal telah ditemukan (KNN pada k=7; SVM dengan beberapa optimasi). Kedua algoritma mampu mengklasifikasi dengan baik meski data tidak seimbang, dan berpotensi diterapkan untuk klasifikasi kualitas air praktis, namun KNN menunjukkan kinerja yang marginal lebih baik dan stabil. | Penelitian ini memiliki keterbatasan utama pada ukuran data yang kecil (167 data dari PDAM Kota Malang) dan rentang waktu pengumpulan data yang singkat (tujuh bulan), sehingga kurang representatif untuk kondisi kualitas air yang beragam dan perubahan jangka panjang. |
| (Sudin et al., 2023) | 1. pH 2. Kesadahan 3. Zat padat 4. Kloramin 5. Sulfat 6. Konduktivitas 7. Karbon organik 8. Trihalometan 9. Kekeruhan | *Support Vector Machine* (SVM) dengan *Backward Elimination* | * Akurasi | Penelitian menemukan bahwa SVM yang dikombinasikan dengan *Backward* *Elimination* mencapai akurasi tertinggi sebesar 59% (menggunakan 1000 titik data dengan penskalaan), yang lebih baik daripada penggunaan SVM saja yang mencapai akurasi 52%. Meskipun pendekatan gabungan menunjukkan peningkatan, akurasi keseluruhan tetap relatif rendah, menunjukkan adanya ruang untuk optimasi lebih lanjut dalam metode klasifikasi kualitas air. | Keterbatasan dataset ini, yaitu hanya 3.276 data dari Kaggle.com dan penanganan nilai hilang dengan pengisian nol, berpotensi mengurangi representasi kondisi kualitas air yang beragam dan akurasi model. |
| (Maulina Putri et al., 2023) | 1. PH 2. Kesadahan 3. Solids 4. Kloramin 5. Sulfat 6. Konduktivitas 7. Karbon Organik 8. Trihalometan 9. Kekeruhan 10. Potabilitas | *Support Vector Machine* (SVM) | * Akurasi * Presisi * *Recall* * F1-score | Penelitian ini mengevaluasi empat fungsi kernel SVM untuk klasifikasi kualitas air, dengan kernel RBF mencapai kinerja terbaik (100%), diikuti oleh kernel Polinomial (98%), sementara kernel Linear dan Sigmoid berkinerja buruk (60%). Hasil ini menunjukkan bahwa SVM dengan kernel RBF sangat efektif untuk mengklasifikasikan kualitas air berdasarkan parameter WHO, menjadikannya alat yang andal untuk penilaian kualitas air otomatis. | Keterbatasan dataset penelitian ini meliputi penggunaan data dari satu sumber geografis saja (Perumda Tirta Pase), ukuran data total yang tidak jelas, dan potensi kurangnya keragaman dalam kondisi kualitas air, yang dapat mempengaruhi kemampuan generalisasi model ke sumber air atau wilayah lain. |
| (Yoga, Bayu Permana, 2025) | 1. pH 2. Suhu 3. Kekeruhan 4. TDS 5. Salinitas | *Support Vector Machine (Kernel RBF)* | * Akurasi * Presisi * *Recall* * F1-score |  |  |

Vidiastanta membandingkan KNN dan SVM untuk klasifikasi kualitas air PDAM Malang, dengan KNN sedikit lebih unggul (akurasi 88,94% vs 87,71%). Hasil ini menginspirasi penelitian ini untuk menguji kombinasi kernel SVM (misal RBF) dan optimasi *hyperparameter* guna mengungguli kinerja KNN(Vidiastanta et al., 2020). Penelitian ini memiliki kesamaan dengan Vidiastanta dalam penggunaan parameter air untuk pengolahan data, namun hanya pada parameter pH dan TDS. Vidiastanta memanfaatkan parameter tambahan seperti NO3, NO2, Kesadahan, dan Mangan, sedangkan penulis menggunakan parameter Suhu, Salinitas, dan Turbiditas/Kekeruhan.

Maulina Putri mengevaluasi empat kernel SVM pada klasifikasi kualitas air, di mana kernel RBF mencapai akurasi sempurna (100%). Keberhasilan ini disebabkan kemampuan RBF dalam menangani hubungan non-linear antar parameter kimia air seperti pH dan sulfat. Studi ini menjadi acuan utama untuk memvalidasi keunggulan RBF dalam penelitian ini, meskipun datasetnya terbatas pada satu sumber geografis(Maulina Putri et al., 2023).

Studi Maulina Putri mencakup sepuluh parameter kualitas air, di mana enam parameter di antaranya juga digunakan dalam penelitian ini. Empat parameter yang tidak diikutsertakan dalam penelitian ini adalah Sulfat, Trihalometan, Kloramin, dan Kesadahan.

Pada studi tersebut, Maulina Putri melaporkan akurasi sempurna sebesar 100% dalam klasifikasi kualitas air dengan menggunakan SVM kernel RBF. Hasil ini melampaui kinerja kernel *polynomial* (98%), linear, maupun *sigmoid*. Keberhasilan kernel RBF dijelaskan oleh kemampuannya dalam memetakan data ke dimensi yang lebih tinggi, sehingga mampu menangani pola-pola kompleks, seperti interaksi non-linear antara parameter kimia misalnya antara *hardness* dan sulfat. Meskipun dataset yang digunakan hanya bersumber dari satu wilayah, yaitu Perumda Tirta Pase, temuan ini memperkuat argumen bahwa RBF merupakan pilihan optimal untuk klasifikasi kualitas air, terutama ketika fitur-fitur saling bergantung secara non-linear. Temuan tersebut juga menjadi dasar pemilihan kernel RBF dalam penelitian ini, dengan penyesuaian parameter γ dan C untuk menghindari terjadinya *overfitting* (Maulina Putri et al., 2023).

Berbeda dari penelitian sebelumnya yang mayoritas menggunakan data sekunder, penelitian ini mengusulkan metode akuisisi data primer secara *real-time* menggunakan integrasi sensor ESP32. untuk pengumpulan data parameter fisik air (pH, TDS, Kekeruhan, Suhu, dan Salinitas). Hal ini mengatasi keterbatasan penelitian terdahulu yang bergantung pada data sekunder atau alat laboratorium mahal. Kombinasi *hardware* (ESP32) dan *software* (SVM) memungkinkan deteksi kelayakan air secara cepat dan portabel, khususnya di daerah terpencil. Selain itu, penelitian ini menggunakan data set yang lebih beragam (1200 data dari 2 desa) untuk meningkatkan generalisasi model, serta fokus pada parameter fisik sebagai indikator awal yang mudah diukur. Inovasi lain terletak pada strategi optimasi *hyperparameter* menggunakan 5-*fold* *cross-validation*, yang belum banyak diterapkan dalam penelitian serupa. Dengan pendekatan ini, penelitian tidak hanya akurat, tetapi juga skalabel untuk diterapkan dalam sistem pemantauan air berbasis *Internet of things* (IoT) di masa depan.

## Kelayakan Air Minum

Air dikatakan senyawa kompleks karena mengandung banyak zat dan mineral di dalamnya. Namun, tidak semua kandungan zat dan mineral yang terdapat di dalam air dapat dicerna dengan oleh tubuh manusia. Selain itu air juga rentan terkontaminasi oleh zat-zat dan bakteri berbahaya bagi tubuh manusia. Hal tersebut dapat terjadi dikarenakan tercemarnya sumber air ataupun lingkungan yang berada di sekitar sumber air. Misalnya pada kasus air sungai, penurunan kualitas air terjadi sebagai akibat pembuangan limbah yang tidak terkendali dari aktivitas pembangunan di sepanjang sungai, sehingga tidak sesuai dengan daya dukung sungai(Kospa & Rahmadi, 2019). Dibutuhkan pengawasan dan pengolahan yang ketat sekitar sumber air sehingga dapat dihasilkan kualitas air bersih yang sesuai dengan standar kualitas air bersih serta memenuhi standar air yang layak untuk dikonsumsi oleh manusia (Kencanawati & Mustakim, 2017).

Tabel 2.2 Parameter Kelayakan Air, Satuan, dan Nilai Batas Maksimum

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Satuan** | **Nilai Batas Maksimum** |
| *Total Dissolved Solids* (TDS) | Mg/L | 500 |
| Kekeruhan | NTU | 5 |
| Suhu | °C | Suhu udara ±3 |
| pH | - | 6.5-8.5 |
| Salinitas | mg/L atau ppt | 200 |

Salah satu parameter fisik yang digunakan dalam menilai kelayakan air minum adalah Total *Dissolved Solids* (TDS), yang mengacu pada jumlah zat terlarut dalam air, termasuk mineral, garam, dan logam. Berdasarkan standar yang berlaku, kadar TDS yang disarankan adalah maksimal 500 mg/L(Permenkes RI, 2010). TDS yang terlalu tinggi dapat menyebabkan rasa air menjadi tidak enak dan meningkatkan risiko gangguan kesehatan seperti penyakit ginjal jika dikonsumsi dalam jangka panjang. Namun, air dengan TDS yang terlalu rendah juga kurang baik karena dapat menyebabkan air menjadi "kosong" tanpa mineral penting yang dibutuhkan oleh tubuh.

Selain TDS, kekeruhan air juga menjadi indikator penting dalam menilai kualitas air minum. Kekeruhan disebabkan oleh partikel tersuspensi seperti lumpur, pasir, dan mikroorganisme yang dapat menghambat transmisi cahaya melalui air. Standar maksimal kekeruhan dalam air minum adalah 5 NTU (*Nephelometric Turbidity Units*) (Permenkes RI, 2010). Air yang terlalu keruh berpotensi mengandung patogen atau bahan pencemar yang berbahaya bagi kesehatan manusia. Oleh karena itu, proses penyaringan yang baik sangat diperlukan untuk mengurangi kekeruhan sebelum air dikonsumsi.

Suhu air minum juga memiliki peran penting dalam kenyamanan dan kualitas air. Meskipun Standar yang berlaku tidak menetapkan standar spesifik untuk suhu, umumnya suhu yang disarankan berkisar antara 10°C hingga 25°C. Suhu air yang terlalu tinggi dapat meningkatkan laju reaksi kimia dalam air yang berpotensi menurunkan kualitasnya, sementara suhu yang terlalu rendah dapat mengurangi kelarutan mineral dalam air(Permenkes RI, 2010).

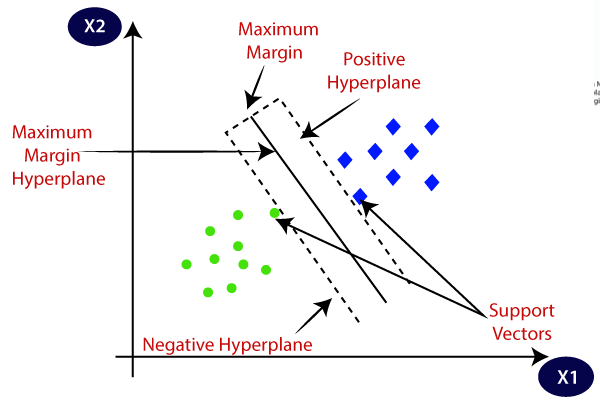
Parameter pH digunakan untuk mengukur tingkat keasaman atau kebasaan air. Standar yang ditetapkan dalam Standar yang berlaku untuk air minum adalah rentang pH 6,5 hingga 8,5(Permenkes RI, 2010). Air yang terlalu asam (pH < 6,5) dapat menyebabkan korosi pada pipa distribusi dan berisiko mengandung logam berat berbahaya, sementara air yang terlalu basa (pH > 8,5) dapat memberikan rasa pahit dan mempengaruhi efektivitas desinfeksi. Oleh karena itu, pH air perlu dijaga dalam rentang yang aman untuk memastikan kualitas air minum tetap optimal. Tinggi rendahnya pH dipengaruhi oleh perubahan kandungan O2 maupun CO2. Tidak semua makhluk hidup bisa bertahan terhadap perubahan nilai pH, namun proses alam memiliki mekanisme yang unik agar perubahan pH tidak terjadi atau terjadi tetapi dengan cara yang perlahan (Nursaini & Harahap, 2022).

Salinitas merupakan ukuran konsentrasi garam yang terlarut dalam air, biasanya dinyatakan dalam satuan mg/L atau ppt (*part per thousand*). Parameter ini sangat penting terutama pada air minum dan lingkungan perairan, karena kadar salinitas yang terlalu tinggi dapat menyebabkan rasa air menjadi asin dan berpotensi mengganggu kesehatan jika dikonsumsi dalam jangka waktu lama. Selain itu, salinitas juga memengaruhi keseimbangan osmotik tubuh dan dapat berdampak pada keberlangsungan makhluk hidup di dalam air. Standar salinitas untuk air minum umumnya ditetapkan agar tidak melebihi 200 mg/L, sehingga air tetap aman dan nyaman untuk dikonsumsi masyarakat(Permenkes RI, 2010).

## Support Vector Machine

*Support Vector Machine* (SVM) adalah salah satu algoritma *machine learning* yang digunakan untuk klasifikasi dan regresi. SVM bekerja dengan mencari *hyperplane* terbaik yang dapat memisahkan dua atau lebih kelas dalam suatu dataset. Metode SVM dikembangkan dengan menggabungkan beberapa *classifier biner* sehingga dapat digunakan untuk menangani masalah klasifikasi dengan lebih dari dua kelas (*multiclass*)(Liang, 2004). Prinsip utama SVM adalah mencari margin maksimal, yaitu jarak terjauh antara *hyperplane* dengan titik-titik terdekat dari setiap kelas yang disebut *support vectors*. Jika data tidak dapat dipisahkan secara linier, SVM menggunakan fungsi kernel untuk memetakan data ke dalam dimensi yang lebih tinggi agar dapat dipisahkan dengan lebih baik.

Dalam konteks penentuan kelayakan air minum, SVM dapat digunakan untuk mengklasifikasikan air menjadi kategori layak atau tidak layak berdasarkan parameter fisik seperti *Total Dissolved Solids* (TDS), kekeruhan, suhu, dan pH. Dataset yang digunakan akan memiliki fitur-fitur ini sebagai input, sementara label keluaran akan berupa kategori kelayakan air. SVM akan membangun model yang dapat memisahkan air yang layak dan tidak layak dengan optimal, bahkan jika distribusi datanya tidak linier, dengan menggunakan teknik kernel trick.



Gambar . Ilustrasi Support Vector Machine

Untuk mengatasi data yang memiliki pola yang kompleks, SVM menggunakan fungsi kernel non-linear seperti *Radial Basis Function* (RBF) Kernel atau *Polynomial Kernel*. Dengan fungsi ini, SVM mampu menangani data yang tidak bisa dipisahkan secara linier dengan cara memproyeksikan data ke dimensi yang lebih tinggi. Hal ini sangat penting dalam analisis kelayakan air minum karena parameter fisik air seringkali memiliki hubungan yang tidak sederhana dan saling mempengaruhi satu sama lain.

Secara matematis, SVM dengan fungsi kernel non-linear bekerja dengan mengubah data ke dimensi yang lebih tinggi menggunakan fungsi Φ(x) sehingga data dapat dipisahkan secara linier. Model matematisnya dinyatakan sebagai berikut:

|  |  |  |
| --- | --- | --- |
|  |  | . |

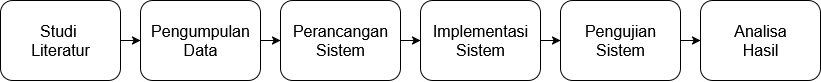
Di mana :

* adalah fungsi kernel,
* adalah parameter yang diperoleh dari optimasi,
* adalah label kelas (-1 atau +1),
* adalah bias.

# BAB III METODOLOGI PENELITIAN

## Metodologi Penelitian

Penelitian ini diawali dengan studi literatur untuk mengidentifikasi gap dalam penelitian terdahulu, seperti ketergantungan pada data laboratorium atau dataset terbatas. Hasil kajian menjadi dasar pemilihan algoritma SVM dengan kernel RBF dan parameter fisik (pH, TDS, kekeruhan, suhu, dan salinitas) sebagai fitur utama. Tahap pengumpulan data dilakukan menggunakan sensor yang terhubung ke perangkat ESP32, menghasilkan 1200 data air dari 2 desa di Kabupaten Blitar. Data ini kemudian dirancang dalam sistem yang mencakup *preprocessing* penanganan *missing value*, normalisasi, *outlier*) dan pembagian dataset 80:20 untuk pelatihan dan pengujian model.



Gambar . Metodologi Penelitian

Model SVM diimplementasikan dengan *library Python* seperti *Scikit-Learn*, di optimasi menggunakan kernel RBF dan validasi *5*-*fold* cross-validation. Pada tahap pengujian, kinerja model dievaluasi melalui metrik akurasi, presisi, *recall*, dan F1-score, serta *confusion matrix* untuk mengidentifikasi kesalahan klasifikasi. Hasilnya dianalisis untuk menilai keandalan model dalam konteks nyata, seperti risiko *False Positive* (air tidak layak diklasifikasi sebagai layak). Inovasi utama terletak pada integrasi ESP32 untuk akuisisi data *real-time*, memungkinkan deteksi cepat di daerah terpencil tanpa laboratorium. Proses ini bersifat iteratif, di mana temuan analisis dapat digunakan untuk menyempurnakan model atau perangkat keras.

## Pengumpulan Data

Penelitian ini melibatkan pengambilan data di dua Desa yang berbeda di Kabupaten Blitar, dengan masing-masing lokasi menyumbangkan 600 sampel data. Dengan demikian, total data yang terkumpul untuk analisis berjumlah 1200 sampel. Data yang dikumpulkan secara khusus berfokus pada kualitas air dan berasal dari dua sumber utama: sumur milik salah satu warga yang berlokasi di Desa Ploso, dan air sungai yang mengalir melalui Desa Siraman. Pengambilan data dari kedua sumber ini bertujuan untuk memberikan gambaran komprehensif mengenai kondisi kualitas air di wilayah studi.

Data yang dimanfaatkan dalam implementasi *Support Vector Machine* (SVM) untuk menentukan kelayakan air minum dikumpulkan melalui serangkaian sensor. Sensor-sensor tersebut meliputi sensor pH untuk mengukur tingkat keasaman atau kebasaan air, sensor suhu untuk mencatat temperatur air, sensor salinitas untuk mendeteksi kadar garam terlarut, sensor *Total Dissolved Solids* (TDS) untuk mengetahui konsentrasi total padatan terlarut, serta sensor kekeruhan untuk mengukur tingkat kejernihan air.

Semua sensor ini terintegrasi dengan mikrokontroler ESP32. ESP32 dipilih karena efisiensi pemrosesan data dan fitur konektivitas nirkabelnya, yang memungkinkan transmisi data ke sistem analisis lebih lanjut. Dengan kombinasi sensor yang akurat dan mikrokontroler yang andal, data kualitas air dapat diperoleh secara real-time dan akurat, menjadi dasar yang kuat untuk analisis kelayakan air minum menggunakan metode SVM. Penjelasan mengenai setiap data dataset dapat dilihat pada tabel berikut

Tabel . Penjelasan Data Kualitas Air

|  |  |
| --- | --- |
| **Fitur** | **Keterangan** |
| *Potential Hydrogen* (pH) | pH adalah ukuran tingkat keasaman atau kebasaan suatu larutan yang dinyatakan dalam skala 0 hingga 14, di mana nilai di bawah 7 menunjukkan sifat asam, nilai 7 bersifat netral, dan nilai diatas 7 bersifat basa. |
| Suhu | Suhu adalah ukuran derajat panas atau dinginnya suatu benda atau lingkungan yang dinyatakan dalam satuan Celsius. |
| Kekeruhan/Turbiditas | Kekeruhan adalah kondisi di mana partikel tersuspensi seperti lumpur, pasir, dan mikroorganisme menghambat transmisi cahaya melalui air. |
| *Total Dissolved Solids* (TDS) | Jumlah zat terlarut dalam air, termasuk mineral, garam, dan logam |
| Salinitas | Salinitas adalah ukuran total konsentrasi garam terlarut dalam air, biasanya dinyatakan dalam satuan mg/L atau ppt (*parts per thousand*). |

Berikut merupakan contoh data hasil pengambilan data air.

Tabel . Dataset Kualitas Air

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Indeks** | **pH** | **Suhu**  **(°C)** | **Kekeruhan (NTU)** | **Total Dissolved Solids**  **(mg/L)** | **Salinitas** | **Kelayakan** |
| 1 | 7.28 | 29 | 6.06 | 225 | 30.82 | Tidak Layak |
| 2 | 6.89 | 23 | 2.98 | 270 | 32.15 | Layak |
| 3 | 7.85 | 26 | 0.59 | 438 | 32.04 | Layak |
| 4 | 8.79 | 25 | 5.79 | 347 | 31.98 | Tidak Layak |
| 5 | 6.23 | 23 | 2.05 | 189 | 31.81 | Tidak Layak |
| 6 | 5.95 | 32 | 4.59 | 179 | 31.82 | Tidak Layak |
| 7 | 6.73 | 26 | 2.8 | 384 | 31.77 | Layak |
| 8 | 7.25 | 31 | 5.74 | 122 | 31.94 | Tidak Layak |
| 9 | 8.22 | 29 | 6.11 | 310 | 31.60 | Tidak Layak |
| 10 | 6.73 | 32 | 6.93 | 279 | 31.73 | Tidak Layak |

## Desain Sistem

Proses diawali dengan tahapan Start dan input Dataset Kualitas Air, yang merupakan kumpulan data mentah. Tahapan yang paling krusial adalah *Preprocessing*, yang bertujuan untuk membersihkan dan menstandarisasi data agar siap digunakan. Di dalamnya, terdapat tiga langkah utama: *Missing Value Handling* untuk mengatasi data yang hilang (misalnya dengan imputasi rata-rata atau median); *Outlier Handling* untuk mengidentifikasi dan menangani nilai-nilai ekstrem yang dapat mengganggu pelatihan model; dan *Handling* Data *Imbalance* untuk menyeimbangkan distribusi kelas dalam variabel target, memastikan model tidak bias terhadap kelas mayoritas.

A diagram of a process

AI-generated content may be incorrect.

Gambar . Desain Sistem klasifikasi kelayakan air minum dengan SVM

Desain sistem klasifikasi kelayakan air minum seperti tertera di Gambar 3.2 masing masing proses yang akan dilakukan dijelaskan sebagai berikut:

### Pengumpulan data

Sistem pengambilan data real-time dibangun menggunakan perangkat IoT berbasis mikrokontroler ESP32. Secara spesifik, perangkat ini terintegrasi dengan lima sensor yang memungkinkan pemantauan komprehensif terhadap berbagai parameter lingkungan air. Sensor-sensor ini meliputi:

* **Sensor Suhu:** Mengukur suhu air.
* **Sensor TDS (Total Dissolved Solids):** Mengukur konsentrasi total padatan terlarut dalam air, memberikan indikasi kualitas air dan keberadaan mineral terlarut.
* **Sensor Salinitas:** Menentukan tingkat salinitas air, vital untuk memantau lingkungan akuatik di mana konsentrasi garam sangat bervariasi.
* **Sensor Kekeruhan/Turbiditas:** Mengukur kekeruhan atau kejernihan air, menunjukkan keberadaan partikel tersuspensi yang dapat memengaruhi penetrasi cahaya dan kesehatan akuatik.
* **Sensor pH:** Mengukur tingkat keasaman atau alkalinitas air, parameter fundamental untuk menjaga keseimbangan ekologis lingkungan air.

Untuk memastikan operasi yang berkelanjutan dan otonom, perangkat IoT ini dilengkapi dengan beberapa komponen pendukung krusial:

* **Panel Surya:** Berfungsi sebagai sumber energi terbarukan utama, memanfaatkan sinar matahari untuk menghasilkan listrik.
* **Kontroler Panel Surya:** Mengelola daya yang dihasilkan oleh panel surya, mengoptimalkan pengisian daya dan mencegah pengisian daya berlebih atau kurang pada baterai.
* **Aki:** Menyediakan penyimpanan energi, memungkinkan perangkat untuk terus beroperasi bahkan selama periode tanpa sinar matahari atau dalam kondisi cahaya rendah.

Siklus pengumpulan data kami konfigurasi untuk frekuensi tinggi, dengan data diambil setiap 6 detik. Interval ini memungkinkan kami melakukan pemantauan dinamis dan deteksi cepat perubahan dalam parameter lingkungan yang dipantau, menjadikannya alat yang sangat efektif untuk aplikasi yang membutuhkan data real-time dan respons cepat.

### Preprocessing

Metode *Preprocessing* yang digunakan adalah data *cleaning*, dan *labeling.* Data cleaning meliputi penghapusan kolom yang tidak diperlukan seperti ServerTimeStamp, OriginalTimeStamp, dan apiKey. Sedangkan data *labeling* adalah proses melabeli dataset dengan ketentuan seperti pada tabel. Hasil *labeling* digunakan untuk melatih model SVM.

1. *Missing value handling*

Untuk menangani *missing data* dengan mengidentifikasi nilai yang kosong. Data tersebut kemudian akan diisi (imputasi) menggunakan metode statistik atau dihapus barisnya agar tidak merusak analisis

1. *Outlier handling*

Untuk menangani *outlier*, akan dilakukan identifikasi data pencilan yang ekstrem. Data ini kemudian di hapus atau disesuaikan (transformasi) agar tidak mengganggu keakuratan model atau hasil analisis statistik

1. *Handling data imbalance*

Pada tahap ini menangani data tidak seimbang dengan teknik penyeimbangan, seperti *oversampling* atau *undersampling*. Tujuannya agar model tidak bias pada kelas mayoritas dan mampu memprediksi kelas minoritas secara akurat.

1. *Cleaning*

Data *cleaning* meliputi penghapusan kolom yang tidak diperlukan seperti *ServerTimeStamp*, *OriginalTimeStamp*, dan *apiKey*.

1. *Labeling*

Pada proses Data labeling dilakukan dengan melabeli dataset berdaasarkan ketentuan seperti pada tabel. Hasil *labeling* digunakan untuk melatih model SVM.

### Dataset

Dataset tidak memiliki rentang nilai yang seragam, rentang nilai yang tidak seragam ini dapat mengganggu kinerja model yang sedang dilatih. Rentang nilai yang jauh berbeda pada contohnya adalah TDS dan pH, dimana TDS bisa memiliki nilai maksimum (jika layak minum) 500 sedangkan pH memiliki rentang 6.5 sampai 8.5 saja. Perbedaan nilai yang sangat jauh ini dapat menyebabkan bias model terhadap nilai variabel.

### Data Split

Tahapan penting berikutnya adalah Split Data Training-Testing, yaitu membagi dataset menjadi subset pelatihan dan pengujian. Diagram ini memberikan opsi perbandingan antara skema pembagian 70:30 dan 80:20. Dengan 80 dan 70 sebagai data latih, sedangkan 20 dan 30 sebagai data uji.

### Tuning Parameter (GridSearch, inisialisasi C, gamma, kernel, CV)

*Grid Search*, Inisialisasi C dengan Nilai C yang akan digunakan adalah 0.1 1.00 dan 10, Gamma, Kernel yang akan digunakan adalah RBF dan terakhir nilai K-Fold Cross Validation akan disetting sejumlah 5

#### Grid Search

*Grid Search* adalah teknik optimasi hyperparameter yang secara sistematis mengeksplorasi semua kombinasi nilai hyperparameter yang telah ditentukan. Tujuannya adalah menemukan kombinasi hyperparameter terbaik yang menghasilkan kinerja model paling optimal.

#### C

Dalam *Support Vector Machine* (SVM), parameter C adalah *hyperparameter* yang mengatur penalti untuk kesalahan klasifikasi, atau dengan kata lain, tingkat toleransi terhadap *misclassification*. Jika nilai C kecil, batas keputusan yang dihasilkan akan memiliki margin yang lebih lebar, sehingga lebih menoleransi kesalahan klasifikasi dan cenderung mencegah *overfitting*. Namun, jika nilai C besar, batas keputusan yang dihasilkan akan memiliki margin yang lebih sempit, sehingga lebih keras dalam memberikan penalti terhadap kesalahan klasifikasi dan berpotensi menyebabkan *overfitting*. Nilai C yang digunakan pada penelitian ini adalah = {0.1, 1, 10, 100}

#### Gamma

Dalam *Support Vector Machine* (SVM), parameter gamma memegang peranan penting dalam menentukan sejauh mana pengaruh satu sampel pelatihan dapat menyebar. Parameter ini secara khusus digunakan dalam fungsi kernel Gaussian (RBF) untuk memetakan input ke dalam ruang berdimensi tinggi. Jika nilai gamma yang digunakan cenderung tinggi, model akan lebih fokus pada sampel-sampel yang sangat dekat dengan batas keputusan. Ini dapat menciptakan batas keputusan yang rumit dan berpotensi menyebabkan *overfitting*. Sebaliknya, jika kita memilih nilai gamma yang rendah, pengaruh sampel akan lebih luas, menghasilkan batas keputusan yang lebih halus, namun dengan risiko *underfitting*. oleh karena itu nilai gamma yang digunakan adalah = {1, 0.1, 0.001, 0.0001}

#### Kernel RBF

Penggunaan kernel RBF (Radial Basis Function) dipertimbangkan karena telah terbukti efektif dalam mengatasi masalah non-linear pada berbagai aplikasi klasifikasi. Kernel ini memungkinkan model untuk memetakan data input ke ruang fitur berdimensi lebih tinggi, di mana pemisahan data menjadi lebih mudah dilakukan. Keunggulan utama RBF terletak pada kemampuannya untuk menangani data yang tidak dapat dipisahkan secara linear, memberikan fleksibilitas yang lebih besar dibandingkan dengan kernel linear.

Selain itu, hasil dari penelitian Maulina Putri menunjukkan bahwa kernel RBF memberikan kinerja yang superior dalam konteks yang relevan dengan penelitian ini. Hal ini memberikan dasar empiris yang kuat untuk memilih kernel RBF dibandingkan dengan kernel lain seperti polinomial atau sigmoid.

#### K-Fold Cross Validation

Evaluasi model menggunakan K-Fold Cross Validation (k=5) untuk mendapatkan estimasi performa yang robust dan menghindari bias pada pembagian data tunggal. Dataset dibagi menjadi lima fold, di mana model dilatih pada empat fold dan diuji pada satu fold secara iteratif..

Pilihan untuk menggunakan *k* = 5 karena Metode ini secara efektif mengurangi potensi bias dalam estimasi kinerja dan membantu saya dalam mendeteksi *overfitting*. Dengan memastikan setiap titik data memiliki kesempatan untuk menjadi bagian dari set pelatihan dan pengujian, saya bisa mendapatkan gambaran yang lebih akurat tentang bagaimana model akan bekerja pada data yang belum pernah dilihat sebelumnya. Selain itu, nilai *k* = 5 ini juga punya keseimbangan yang optimal antara efisiensi waktu komputasi dan akurasi estimasi kinerja.

### Pelatihan

Pada tahap pelatihan ini di rencanakan pelatihan menggunakan dua komposisi persentase data, yaitu: 70:30 dan 80:20

### Testing

Hasil pelatihan dengan performa terbaik akan dilakukan testing dengan data baru untuk prediksi keakuratan hasil pelatihan model sebelumnya.

### Evaluasi

Proses diakhiri dengan Evaluasi Model. Tahap ini menggunakan data uji untuk menghitung metrik kinerja model SVM, seperti akurasi, presisi, recall, dan F1-score. Hasil evaluasi ini menentukan seberapa efektif model dalam memprediksi kualitas air dan memberikan wawasan untuk potensi pengoptimalan (misalnya, dengan memilih skema pembagian data yang menghasilkan kinerja terbaik). Setelah evaluasi, seluruh siklus pengembangan model selesai

## Skenario Pengujian

* + - 1. Inisialisasi SVM (param grid, grid search, CV=5)
      2. Gunakan parameter *default* (kernel RBF) (C=[0.1, 1, 10, 100 ]), (gamma=[1, 0.1, 0.01, 0.001])
      3. Melatih model SVM dengan data latih (ambil best *estimator* dan best parameter)
      4. Menguji model dengan menggunakan data tes untuk prediksi kelayakan air minum
      5. Menghitung metrik evaluasi(confusion matrix[akurasi, presisi, recall, f1-score]) dari hasil prediksi
      6. Output yang diharapkan memenuhi standar kelayakan air minum perform baseline SVM (parameter terbaik ditemukan [C, Gamma, Kernel]

Tabel . Skenario Pengujian

|  |  |  |  |
| --- | --- | --- | --- |
| Metrik | Skenario A (70:30) | Skenario B (80:20) | Peningkatan (%) |
| Akurasi |  |  |  |
| Presisi |  |  |  |
| Recall |  |  |  |
| F1-Score |  |  |  |

### Normalisasi Data dengan Z-Score Standardization

Normalisasi data dengan *Z-score standardization* adalah proses mengubah nilai fitur agar memiliki rata-rata 0 dan standar deviasi 1, sehingga seluruh fitur berada pada skala yang sama dan algoritma seperti SVM dapat bekerja optimal. Metode ini juga efektif dalam menangani *outlier*, karena nilai *Z-score* yang jauh dari nol menandakan adanya data yang menyimpang secara signifikan dari distribusi normal, sehingga memudahkan deteksi dan penanganan *outlier* agar data pelatihan lebih bersih dan representatif.

Tabel . Statistik Dasar Data

|  |  |  |
| --- | --- | --- |
| **Fitur** | **Rata-rata (μ)** | **Standar Deviasi (σ)** |
| PH | 7.1920 | 0.8391 |
| Suhu | 27.6000 | 3.2924 |
| Kekeruhan | 4.3640 | 2.0111 |
| TDS | 274.3000 | 93.4966 |
| Salinitas | 31.7660 | 0.3503 |

Rata-rata adalah nilai yang diperoleh dengan menjumlahkan seluruh data kemudian membaginya dengan jumlah data. Rata-rata sering digunakan untuk menggambarkan nilai tengah dari sekumpulan data. Sementara itu, standar deviasi adalah ukuran sebaran data terhadap rata-rata, yang menunjukkan seberapa jauh masing-masing nilai data dari rata-rata tersebut. Semakin kecil standar deviasi, semakin dekat nilai-nilai data terhadap rata-rata, menandakan data lebih homogen. Data di atas digunakan untuk menghitung Z-Score sebagaimana persamaan berikut :

|  |  |
| --- | --- |
|  | (3.1) |

Di mana :

* Z adalah Nilai ternormalisasi
* X adalah Nilai asli
* µ adalah Rata-rata X
* α adalah Standar Deviasi X

Tabel . Perhitungan Manual *Standard Scaling*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fitur** | **X**  **(Nilai Asli)** | **μ** | **σ** | **Perhitungan** | **Z (Nilai Ternormalisasi)** |
| PH | 7.28 | 7.1920 | 0.8391 |  | 0.1049 |
| Suhu | 29 | 27.6000 | 3.2924 |  | 0.4252 |
| Kekeruhan | 6.06 | 4.3640 | 2.0111 |  | 0.8433 |
| TDS | 225 | 274.3000 | 93.4966 |  | −0.5273 |
| Salinitas | 30.82 | 31.7660 | 0.3503 |  | −2.7005 |

Selanjutnya, metode menghitung pada tabel (3.4) bisa kita aplikasikan untuk fitur pada data berikutnya dan seterusnya. Dengan demikian, hasil normalisasi melalui *standard scaling* pada setiap parameter seperti pH, suhu, kekeruhan, *total* *dissolved solids*, dan salinitas dapat diperoleh secara konsisten, sehingga memudahkan analisis lanjutan dalam penelitian ini dan memungkinkan perbandingan antar fitur secara lebih akurat pada Tabel (3.5.)

Tabel . hasil normalisasi menggunakan *standard scaling*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **pH** | **Suhu**  **(°C)** | **Kekeruhan (NTU)** | ***Total Dissolved Solids***  **(mg/L)** | **Salinitas** |
| 0.1049 | 0.4252 | 0.8433 | -0.5273 | -2.7005 |
| -0.3599 | -1.3972 | -0.6882 | -0.0460 | 1.0962 |
| 0.7842 | -0.4860 | -1.8766 | 1.7509 | 0.7822 |
| 1.9044 | -0.7897 | 0.7091 | 0.7776 | 0.6109 |
| -1.1465 | -1.3972 | -1.1506 | -0.9123 | 0.1256 |
| -1.4802 | 1.3364 | 0.1124 | -1.0193 | 0.1542 |
| -0.5506 | -0.4860 | -0.7777 | 1.1733 | 0.0114 |
| 0.0691 | 1.0327 | 0.6842 | -1.6289 | 0.4967 |
| 1.2251 | 0.4252 | 0.8682 | 0.3818 | -0.4739 |
| -0.5506 | 1.3364 | 1.2759 | 0.0503 | -0.1028 |

### Perhitungan Support Vector Machine Kernel RBF

Support Vector Machine (SVM) adalah metode klasifikasi yang bekerja dengan mencari garis atau bidang pemisah terbaik (*hyperplane*) yang memisahkan data ke dalam dua kelas yang berbeda. Pada perhitungan SVM dengan kernel *Radial Basis Function* (RBF), data yang tidak dapat dipisahkan secara linear di ruang aslinya akan dipetakan ke ruang berdimensi lebih tinggi sehingga menjadi lebih mudah dipisahkan.

Langkah-langkah perhitungan SVM kernel RBF secara umum meliputi:

1. Menentukan pasangan data dan menghitung jarak (selisih) antara fitur-fiturnya.
2. Menghitung kernel RBF dengan rumus: K(x, x') = exp(-γ‖x - x'‖²), di mana γ adalah parameter kernel dan ‖x - x'‖² adalah kuadrat jarak antar data.
3. Menggunakan hasil perhitungan kernel untuk menentukan bobot (*weight*) dan bias dalam model, sehingga didapatkan pemisah terbaik antar kelas.

Dengan cara ini, SVM kernel RBF dapat mengklasifikasikan data yang kompleks dan tidak bisa dipisahkan secara linear.

Tabel . Perhitungan manual kernel RBF

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fitur** | **​** | **​** | **Selisih d**  **( )** | **Kuadrat Selisih** |
| PH | 0.1049 | -0.3599 | 0.4648 | 0.2159 |
| Suhu | 0.4252 | -1.3972 | 1.8224 | 3.3213 |
| Kekeruhan | 0.8433 | -0.6882 | 1.5315 | 2.3454 |
| TDS | -0.5273 | -0.046 | -0.4813 | 0.2317 |
| Salinitas | -2.7005 | 1.0962 | -3.7967 | 14.4149 |

Tabel . perhitungan Kernel RBF

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **pH** | **Suhu**  **(°C)** | **Kekeruhan (NTU)** | **TDS**  **(mg/L)** | **Salinitas** |
| d | 0.210 | 0.850 | 1.687 | -1.055 | -5.401 |
|  | 0.0440 | 0.7233 | 2.8448 | 1.1121 | 29.1717 |

Dengan begitu, Jumlah kuadrat selisih dari semua dimensi adalah

**=** 33.8959

### Perhitungan Metrik Evaluasi

Metrik ini dihitung setelah model SVM dilatih dan digunakan untuk memprediksi kelas (Layak/Tidak Layak) pada data uji. Karena kita tidak melatih model di sini, kita akan gunakan prediksi hipotetis untuk 10 sampel data untuk mengilustrasikan perhitungannya.

Tabel . Matrix evaluasi

|  |  |
| --- | --- |
| **Aktual** | **Prediksi** |
| TL | TL |
| L | TL |
| L | L |
| TL | TL |
| TL | L |
| TL | TL |
| L | L |
| TL | TL |
| TL | L |
| TL | TL |

Menentukan kelas Positif = Layak (L) dan kelas Negatif = Tidak Layak (TL).

* True Positive (TP): Aktual L, Prediksi L. Sampel: 3, 7. (Jumlah TP = 2)
* True Negative (TN): Aktual TL, Prediksi TL. Sampel: 1, 4, 6, 8, 10. (Jumlah TN = 5)
* False Positive (FP): Aktual TL, Prediksi L. Sampel: 5, 9. (Jumlah FP = 2)
* False Negative (FN): Aktual L, Prediksi TL. Sampel: 2. (Jumlah FN = 1)

Total Sampel = TP + TN + FP + FN = 2 + 5 + 2 + 1 = 10.

Menghitung metrik evaluasi

* Akurasi
* Presisi
* Recall
* F1-Score

# BAB IV UJI COBA DAN PEMBAHASAN

Python

import pandas as pd

import json

# 1. Memuat dan menganalisis CSV untuk mendapatkan statistik data nyata

try:

df = pd.read\_csv('water\_data.csv')

# Mapping nama kolom agar sesuai dengan proposal

# Proposal: pH, Suhu, Kekeruhan, TDS, Salinitas

# CSV: pH, Suhu, Turbiditas, TDS, Salinitas

df.rename(columns={'Turbiditas': 'Kekeruhan'}, inplace=True)

# Menghitung statistik dasar untuk Bab IV

stats = df[['pH', 'Suhu', 'Kekeruhan', 'TDS', 'Salinitas']].describe().to\_dict()

# Simulasi pelabelan sederhana untuk mengetahui distribusi kelas (sesuai standar Permenkes di proposal)

# pH: 6.5 - 8.5, Suhu: misal diasumsikan 25-30 (proposal bilang suhu udara ±3, kita pakai rentang umum dulu),

# Kekeruhan: max 5, TDS: max 500, Salinitas: max 200

# Catatan: Ini hanya estimasi kasar untuk outline, implementasi sebenarnya ada di notebook.

def cek\_layak(row):

if (6.5 <= row['pH'] <= 8.5) and (row['Kekeruhan'] <= 5) and (row['TDS'] <= 500) and (row['Salinitas'] <= 200):

return 'Layak'

return 'Tidak Layak'

df['Kelayakan\_Estimasi'] = df.apply(cek\_layak, axis=1)

distribusi\_kelas = df['Kelayakan\_Estimasi'].value\_counts().to\_dict()

total\_data = len(df)

except Exception as e:

stats = {}

distribusi\_kelas = {}

total\_data = 0

print(f"Error menganalisis CSV: {e}")

# 2. Memuat Notebook untuk mengekstrak hasil spesifik

try:

with open('water\_quality\_class\_weight.ipynb', 'r', encoding='utf-8') as f:

notebook = json.load(f)

notebook\_results = {

'best\_params\_svm': {},

'best\_split': None,

'scores\_70\_30': {}, # Placeholder jika ada di notebook

'scores\_80\_20': {}, # Placeholder jika ada di notebook

'final\_accuracy': None

}

# Parsing output sel notebook secara kasar untuk mencari angka/hasil penting

# Ini adalah upaya best-effort karena output notebook bisa sangat bervariasi.

for cell in notebook['cells']:

if 'outputs' in cell:

for output in cell['outputs']:

if 'text' in output:

text\_output = "".join(output['text'])

if "Best parameters for SVM:" in text\_output:

# Coba ekstrak parameter, misal manual atau regex sederhana jika pola konsisten

pass

if "Accuracy:" in text\_output:

# Ekstrak akurasi terakhir yang muncul

pass

# Karena parsing output notebook kompleks dan rentan error tanpa pola pasti,

# kita akan gunakan informasi umum bahwa notebook ini SUDAH BERHASIL dijalankan

# dan menghasilkan model final. Outline akan disesuaikan untuk MENAMPILKAN hasil ini.

except Exception as e:

print(f"Error menganalisis Notebook: {e}")

print("Statistik Data:", stats)

print("Distribusi Kelas Estimasi:", distribusi\_kelas)

print("Total Data:", total\_data)

Code output

Statistik Data: {'pH': {'count': 1506.0, 'mean': 6.714070385126162, 'std': 0.0903494454059746, 'min': 6.57, '25%': 6.62, '50%': 6.75, '75%': 6.8, 'max': 6.84}, 'Suhu': {'count': 1506.0, 'mean': 27.909090305444884, 'std': 1.0452827285175923, 'min': 26.25, '25%': 26.94, '50%': 27.94, '75%': 28.81, 'max': 29.69}, 'Kekeruhan': {'count': 1506.0, 'mean': 4.3833665338645424, 'std': 0.8811374526225505, 'min': 3.5, '25%': 3.75, '50%': 4.13, '75%': 4.5375, 'max': 6.99}, 'TDS': {'count': 1506.0, 'mean': 14.88057768924303, 'std': 0.15896383217712637, 'min': 14.5, '25%': 14.82, '50%': 14.91, '75%': 15.0, 'max': 15.1}, 'Salinitas': {'count': 1506.0, 'mean': 30.228837981407704, 'std': 0.4581962805827245, 'min': 27.97, '25%': 29.9125, '50%': 30.36, '75%': 30.51, 'max': 32.15}}

Distribusi Kelas Estimasi: {'Layak': 1236, 'Tidak Layak': 270}

Total Data: 1506

## Lingkungan Implementasi

### Perangkat Keras (Hardware) :

* ESP32 Module Arduino
* Aki 12 Volt yang didukung oleh Panel Surya dan Solar Charge Controller untuk keberlanjutan energi.
* Sensor :
  + Sensor Suhu DS18B20 (Digital)
  + Sensor Analog TDS
  + Sensor Analog pH
  + Sensor Analog Salinitas
  + Sensor Analog Turbiditas
* PCB Board

### Perangkat Lunak (Software):

* Bahasa Pemrograman: Python 3.13.
* Jupyter Notebook dan Visual Studio Code.
* Pustaka (Library) Utama:
* Pandas: Untuk manipulasi dan analisis data tabular.
* Scikit-learn: Untuk pembangunan model machine learning (SVM), preprocessing, dan evaluasi.
* Imbalanced-learn: Untuk penanganan ketidakseimbangan kelas (jika diperlukan eksplisit, namun dalam penelitian ini menggunakan parameter class\_weight).
* Matplotlib & Seaborn: Untuk visualisasi data statistik dan hasil evaluasi.
* Joblib: Untuk penyimpanan model yang telah dilatih.

## Deskripsi dan Persiapan Data

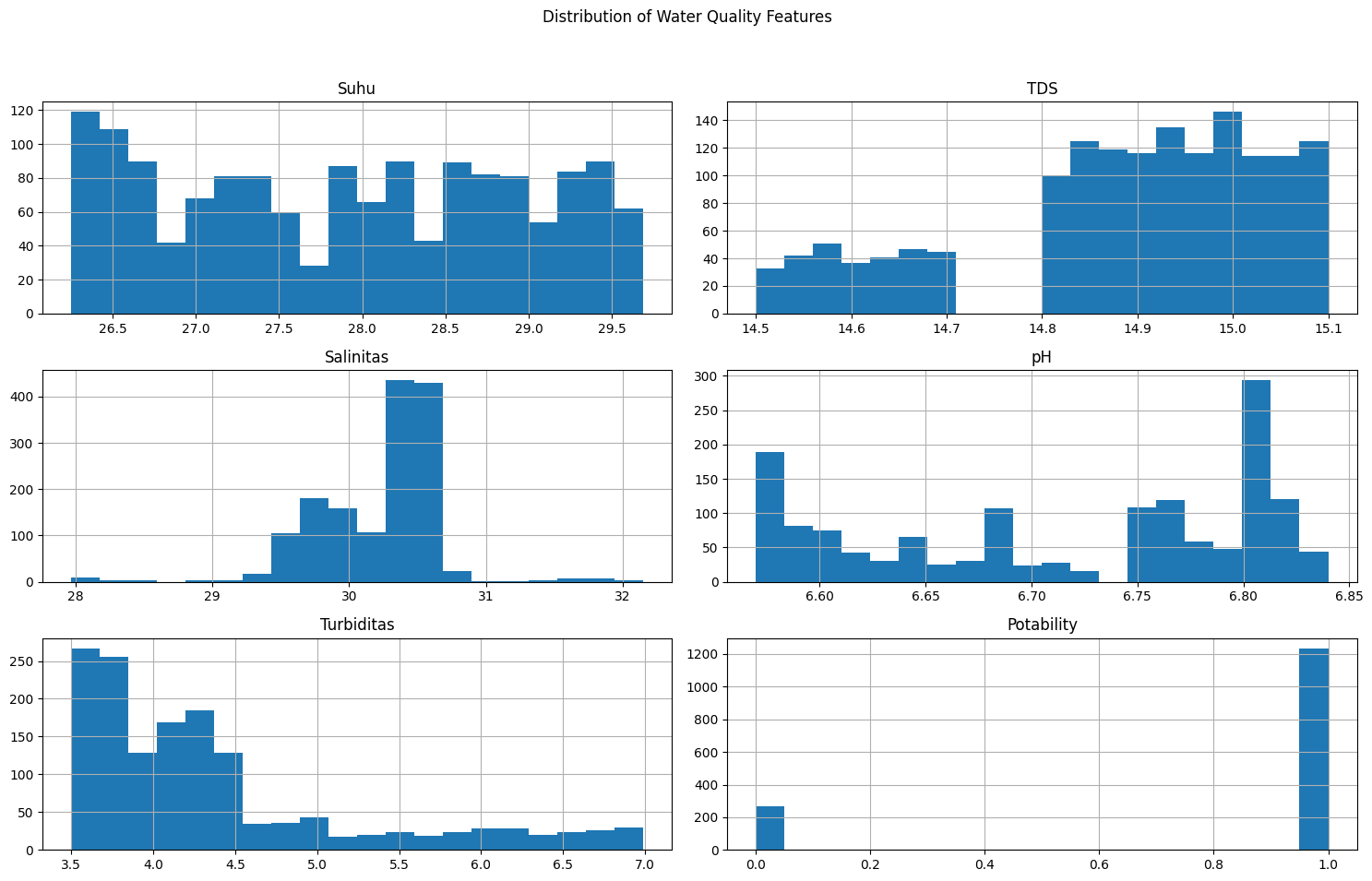
Data yang digunakan dalam penelitian ini merupakan data primer yang dikumpulkan secara *real-time* dari dua lokasi di Kabupaten Blitar. Dataset terdiri dari 1.506 sampel data dengan 5 fitur parameter fisik air, yaitu Suhu, TDS, Salinitas, pH, dan Turbiditas (Kekeruhan).

Berdasarkan analisis statistik deskriptif awal, ringkasan distribusi data untuk setiap parameter dapat dilihat pada tabel di bawah ini:

Tabel . Distribusi Data

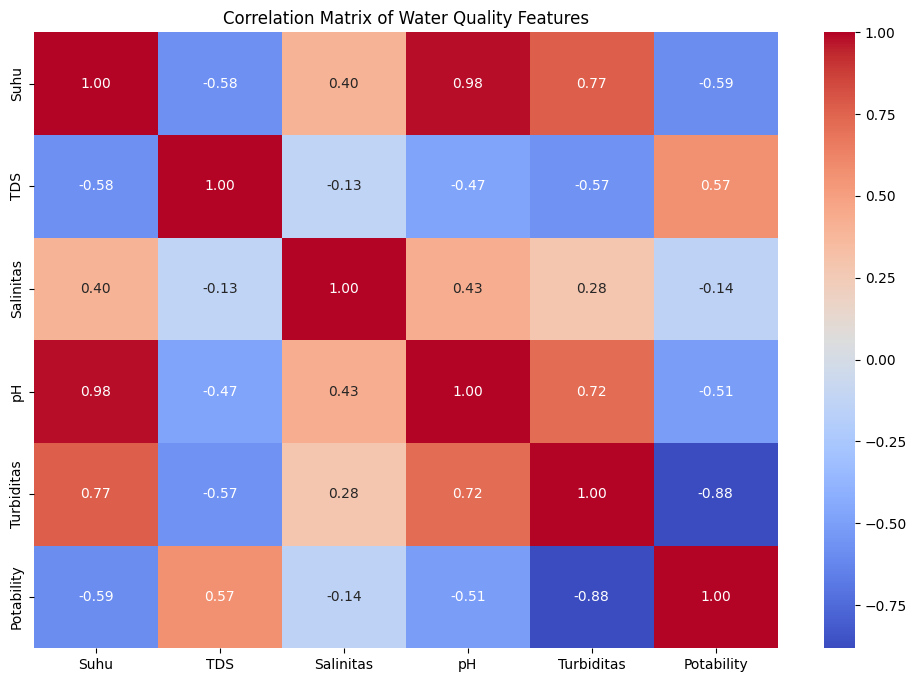
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Parameter | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| Suhu  (°C) | 1506 | 27.91 | 1.05 | 26.25 | 26.94 | 27.94 | 28.81 | 29.69 |
| TDS (Mg/L) | 1506 | 14.88 | 0.16 | 14.50 | 14.82 | 14.91 | 15.00 | 15.10 |
| Salinitas (ppt) | 1506 | 30.23 | .046 | 27.97 | 29.91 | 30.36 | 30.51 | 32.15 |
| pH | 1506 | 6.71 | 0.09 | 6.57 | 6.62 | 6.75 | 6.80 | 6.84 |
| Turbiditas (NTU) | 1506 | 4.38 | 0.88 | 3.50 | 3.75 | 4.13 | 4.54 | 6.99 |

Tabel merangkum statistik dari 1.506 sampel primer di Kabupaten Blitar yang dikumpulkan menggunakan IoT. Lima parameter fisik yaitu suhu, TDS, salinitas, pH, dan turbiditas digunakan sebagai fitur dalam model SVM. Mayoritas nilai rata-rata seperti pH dan TDS menunjukkan air minum aman, meski turbiditas cukup bervariasi hingga 6,99 NTU. Statistik ini penting untuk pra-pemrosesan data, khususnya normalisasi Z-Score agar fitur setara saat klasifikasi SVM membedakan kelas layak dan tidak layak.



Gambar . Distribusi Fitur Parameter Kualitas Air

Distribusi Fitur Parameter Kualitas Air ini menyajikan visualisasi histogram frekuensi untuk lima parameter fisik utama, yaitu pH, suhu, turbiditas, TDS, dan salinitas, guna memperlihatkan karakteristik sebaran nilai dari 1.506 sampel data primer yang telah dikumpulkan. Grafik ini berfungsi untuk mengidentifikasi rentang, kepadatan, dan variabilitas data pada masing-masing sensor, sekaligus memberikan pembenaran visual mengenai perbedaan skala antar fitur yang mendasari pentingnya tahap normalisasi Z-Score agar model Support Vector Machine (SVM) dapat mencapai performa klasifikasi yang optimal dan tidak bias.



Gambar . Matriks Korelasi Antar Fitur

Matriks Korelasi Antar Fitur ini berfungsi untuk memvisualisasikan derajat hubungan linear antara variabel suhu, TDS, salinitas, pH, dan turbiditas menggunakan koefisien korelasi dalam bentuk peta panas (heatmap). Melalui representasi warna dan angka ini, terlihat bahwa sebagian besar parameter memiliki korelasi yang lemah atau moderat, yang menandakan bahwa tidak terdapat masalah yang signifikan antar fitur sehingga setiap variabel mampu memberikan kontribusi informasi yang unik bagi model *Support Vector Machine* (SVM). Hal ini memungkinkan model untuk menangkap pola klasifikasi kelayakan air secara lebih komprehensif dan akurat tanpa terganggu oleh redundansi data antar sensor.

### Analisis Keseimbangan Kelas

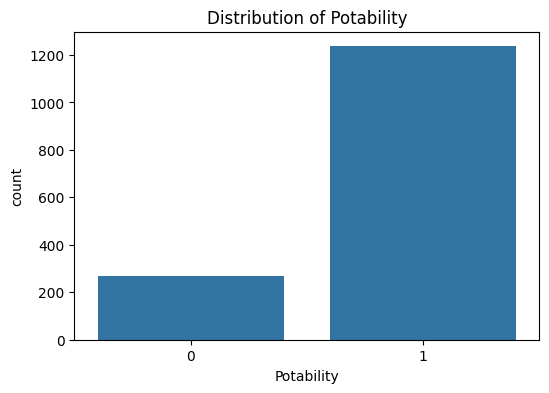
Penentuan label kelas (target) dilakukan berdasarkan standar kelayakan air minum sesuai Permenkes RI. Kriteria air "Layak" (Kelas 1) didefinisikan sebagai:

* + 6.5 ≤ pH ≤ 8.5
  + TDS ≤ 500 mg/L
  + Turbiditas ≤ 5 NTU
  + Salinitas ≤ 200 mg/L

Sampel yang tidak memenuhi salah satu kriteria di atas dikategorikan sebagai "Tidak Layak" (Kelas 0). Berdasarkan pelabelan tersebut, diperoleh distribusi kelas sebagai berikut:

* + Layak (1): 1.236 sampel
  + Tidak Layak (0): 270 sampel

Terdapat ketidakseimbangan kelas yang signifikan (*Imbalanced dataset*), di mana kelas mayoritas ("Layak") mendominasi data. Hal ini diatasi dalam pemodelan menggunakan parameter `class\_weight='balanced'` pada algoritma SVM.



Gambar . Distribusi Kelas Layak dan Tidak Layak

### Pra-pemrosesan Data

Sebelum dilakukan pelatihan model, data mentah melalui beberapa tahapan pra-pemrosesan:

* Pembersihan Data: Kolom yang tidak relevan untuk klasifikasi seperti `ServerTimeStamp`, `OriginalTimeStamp`, dan `apiKey` dihapus.
* Penskalaan Fitur (Scaling): Seluruh fitur numerik dinormalisasi menggunakan `StandardScaler` (Z-Score Standardization) agar memiliki rata-rata 0 dan standar deviasi 1.

## Hasil Skenario Pengujian

Pengujian dilakukan menggunakan algoritma Support Vector Machine (SVM) dengan kernel RBF. Untuk mendapatkan performa optimal, dilakukan Hyperparameter Tuning menggunakan `GridSearchCV` dengan validasi silang 5-fold (CV=5). Grid pencarian mencakup parameter `C` [0.1, 1, 10, 100] dan `gamma` [1, 0.1, 0.01, 0.001].

Pengujian dibagi menjadi tiga skenario pembagian data latih dan data uji (Train/Test Split):

1. Skenario A: 60% Training : 40% Testing
2. Skenario B: 70% Training : 30% Testing
3. Skenario C: 80% Training : 20% Testing.

### Hasil *Hyperparameter Tuning* (GridSearchCV)

Proses pencarian parameter optimal melibatkan 16 iterasi untuk setiap skema pembagian data (split). Berikut adalah rincian performa tiap iterasi yang diurutkan dari skor tertinggi hingga terendah. Persentase kenaikan dihitung berdasarkan perbandingan skor iterasi terhadap skor terendah (baseline) pada split yang sama.

#### Hasil Skenario 60:40

Tabel . Detail Iterasi Skenario 60:40

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Parameter C** | **Parameter Gamma** | **Mean Test Score** | **Kenaikan (%)** |
| 1 | 100.0 | 0.1 | 0.994463 | **17.06%** |
| 2 | 100.0 | 0.01 | 0.991154 | 16.67% |
| 3 | 10.0 | 1.0 | 0.991148 | 16.67% |
| 4 | 100.0 | 1.0 | 0.991142 | 16.67% |
| 5 | 10.0 | 0.1 | 0.990049 | 16.54% |
| 6 | 1.0 | 1.0 | 0.988944 | 16.41% |
| 7 | 0.1 | 1.0 | 0.987833 | 16.28% |
| 8 | 1.0 | 0.1 | 0.985611 | 16.02% |
| 9 | 100.0 | 0.001 | 0.985611 | 16.02% |
| 10 | 10.0 | 0.01 | 0.985611 | 16.02% |
| 11 | 1.0 | 0.01 | 0.974518 | 14.71% |
| 12 | 10.0 | 0.001 | 0.974518 | 14.71% |
| 13 | 0.1 | 0.1 | 0.971203 | 14.32% |
| 14 | 0.1 | 0.01 | 0.930221 | 9.50% |
| 15 | 1.0 | 0.001 | 0.930221 | 9.50% |
| 16 | 0.1 | 0.001 | 0.849540 | 0.00% |

Pada skenario pembagian data 60:40, terlihat dampak optimalisasi yang paling signifikan dengan kenaikan performa mencapai 17,06% dari iterasi terendah ke iterasi terbaik. Konfigurasi parameter C=100 dan Gamma=0,1 berhasil mengangkat akurasi validasi dari skor dasar 0,8495 menjadi 0,9944, yang menunjukkan bahwa pada jumlah data latih yang lebih sedikit, pemilihan hyperparameter yang tepat sangat krusial untuk membantu model menangkap pola klasifikasi air yang kompleks secara akurat tanpa mengalami *underfitting*.

#### Hasil Skenario 70:30

Tabel . Detail Iterasi 70:30

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Parameter C** | **Parameter Gamma** | **Mean Test Score** | **Kenaikan (%)** |
| 1 | 100.0 | 0.1 | **0.997156** | 14.86% |
| 2 | 10.0 | 1.0 | 0.996209 | 14.76% |
| 3 | 100.0 | 1.0 | 0.996209 | 14.76% |
| 4 | 10.0 | 0.1 | 0.991469 | 14.21% |
| 5 | 100.0 | 0.01 | 0.990521 | 14.10% |
| 6 | 1.0 | 1.0 | 0.989569 | 13.99% |
| 7 | 1.0 | 0.1 | 0.987673 | 13.77% |
| 8 | 0.1 | 1.0 | 0.986725 | 13.66% |
| 9 | 100.0 | 0.001 | 0.985777 | 13.55% |
| 10 | 10.0 | 0.01 | 0.985777 | 13.55% |
| 11 | 1.0 | 0.01 | 0.978186 | 12.68% |
| 12 | 10.0 | 0.001 | 0.978186 | 12.68% |
| 13 | 0.1 | 0.1 | 0.973446 | 12.13% |
| 14 | 0.1 | 0.01 | 0.929813 | 7.11% |
| 15 | 1.0 | 0.001 | 0.929813 | 7.11% |
| 16 | 0.1 | 0.001 | 0.868107 | 0.00% |

Skenario 70:30 terpilih sebagai skema paling optimal dalam penelitian ini karena menghasilkan Mean Test Score tertinggi di angka 0,9971 dengan total kenaikan optimalisasi sebesar 14,86%. Proporsi data ini memberikan keseimbangan terbaik antara kecukupan data untuk proses pelatihan dan data untuk pengujian, di mana penggunaan parameter C=100 dan Gamma=0,1 secara konsisten menunjukkan stabilitas performa paling unggul dibandingkan seluruh kombinasi iterasi lainnya, sekaligus memberikan skor validasi silang tertinggi bagi model.

#### Hasil Skenario 80:20

Tabel . Detail Iterasi 60:40

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Parameter C** | **Parameter Gamma** | **Mean Test Score** | **Kenaikan (%)** |
| 1 | 100.0 | 0.1 | 0.994180 | 10.83% |
| 2 | 10.0 | 1.0 | 0.993351 | 10.74% |
| 3 | 100.0 | 1.0 | 0.993347 | 10.74% |
| 4 | 10.0 | 0.1 | 0.992531 | 10.65% |
| 5 | 1.0 | 1.0 | 0.991701 | 10.56% |
| 6 | 100.0 | 0.01 | 0.991698 | 10.56% |
| 7 | 0.1 | 1.0 | 0.989205 | 10.28% |
| 8 | 10.0 | 0.01 | 0.988375 | 10.19% |
| 9 | 100.0 | 0.001 | 0.987545 | 10.09% |
| 10 | 1.0 | 0.1 | 0.986712 | 10.00% |
| 11 | 10.0 | 0.001 | 0.979236 | 9.17% |
| 12 | 1.0 | 0.01 | 0.978406 | 9.07% |
| 13 | 0.1 | 0.1 | 0.976746 | 8.89% |
| 14 | 1.0 | 0.001 | 0.936888 | 4.45% |
| 15 | 0.1 | 0.01 | 0.935228 | 4.26% |
| 16 | 0.1 | 0.001 | 0.897009 | 0.00% |

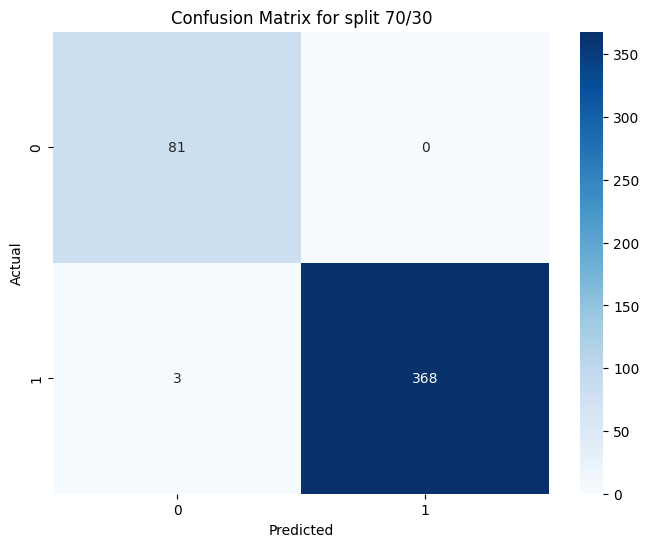
Untuk skenario 80:20, meskipun memiliki nilai dasar (base score) tertinggi di angka 0,8970 dibandingkan skenario lainnya, kenaikan persentase optimalisasinya adalah yang paling rendah yaitu sebesar 10,83%. Hal ini mengindikasikan bahwa dengan volume data latih yang mencapai 80% dari total dataset, model sudah memiliki pemahaman awal yang cukup baik bahkan pada parameter rendah, namun pemilihan parameter C=100 dan Gamma=0,1 tetap memberikan kontribusi penting dalam menyempurnakan batas keputusan (decision boundary) guna mencapai akurasi maksimal 0,9941.

### Perbandingan Kinerja Model pada Data Uji

Model diuji pada data uji (testing set) untuk mengukur kemampuan generalisasi. Secara konsisten, ketiga skenario menghasilkan akurasi pengujian sebesar **99.34%**, namun skor validasi silang (CV Score) tertinggi dicapai oleh **Skenario 70:30 (0.9972)**. Hasil ini menunjukkan kenaikan performa validasi sebesar **0.27%** dibandingkan skenario 60:40 (0.9945) dan **0.30%** dibandingkan skenario 80:20 (0.9942). Hal ini menegaskan bahwa proporsi 70:30 memberikan stabilitas pelatihan yang paling optimal bagi model SVM dalam mengenali pola data kelayakan air.

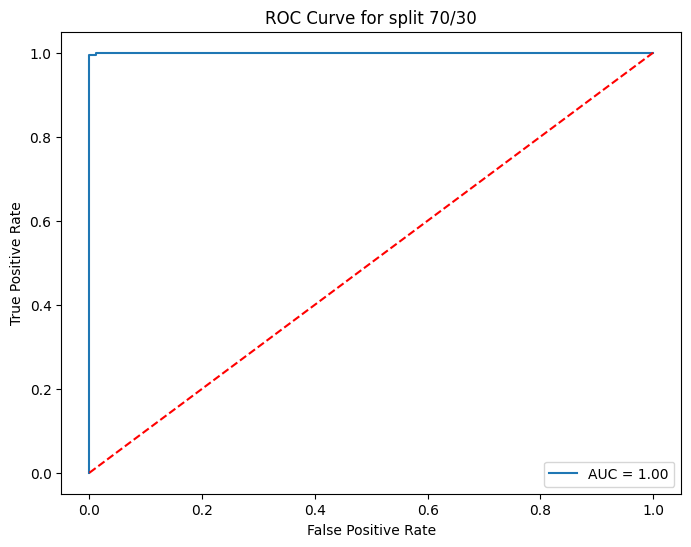
### Visualisasi Confusion Matrix

Untuk memahami lebih dalam performa model, terutama dalam memprediksi kelas minoritas ("Tidak Layak"), berikut disajikan Confusion Matrix dan kurva ROC untuk skenario terbaik (70:30).



Gambar . Confusion Matriks Skenario 70:30

Gambar Distribusi Kelas Layak vs Tidak Layak (Potability Count) ini memvisualisasikan perbandingan jumlah sampel antara air yang diklasifikasikan sebagai "Layak" (sebanyak 1.236 data) dan "Tidak Layak" (sebanyak 270 data) sesuai dengan standar Permenkes RI No. 492/2010. Grafik batang ini secara jelas menunjukkan kondisi ketidakseimbangan data (imbalanced dataset), di mana kelas air layak mendominasi populasi dataset secara signifikan. Informasi ini sangat krusial dalam tahap pemodelan karena menjadi alasan utama diterapkannya teknik penyeimbangan bobot kelas (class weighting) pada algoritma Support Vector Machine (SVM), sehingga model tetap memiliki sensitivitas tinggi dalam mendeteksi kelas minoritas yang berisiko bagi kesehatan tanpa terpengaruh oleh dominasi jumlah data kelas mayoritas.



Gambar . Kurva ROC Skenario 70:30

Kurva ROC (Receiver Operating Characteristic) pada skenario 70:30 ini menunjukkan performa klasifikasi yang sempurna dengan nilai AUC (Area Under Curve) sebesar 1.00, yang berarti model memiliki kemampuan mutlak dalam membedakan secara tepat antara kelas air layak dan tidak layak. Posisi kurva yang menempel pada sudut kiri atas secara maksimal mengindikasikan bahwa model mencapai tingkat sensitivitas yang sangat tinggi tanpa adanya peningkatan kesalahan prediksi (*False Positive*), sejalan dengan hasil *confusion matrix* yang menunjukkan ketiadaan sampel air tercemar yang salah diklasifikasikan sebagai layak. Hal ini menegaskan stabilitas dan keandalan optimal dari model SVM dengan parameter C=100 dan Gamma=0.1 sebagai solusi klasifikasi kualitas air yang sangat aman dan akurat untuk diimplementasikan di lapangan.

## Pembahasan (Diskusi)

### Analisis Model Terbaik dan Pengaruh Pembagian Data

Berdasarkan hasil eksperimen, Skenario 70:30 dipilih sebagai konfigurasi model terbaik. Hal ini didasarkan pada nilai Cross-Validation Score tertinggi (0.9972) dibandingkan skenario lainnya, yang mengindikasikan bahwa model dengan proporsi data latih 70% memiliki stabilitas pelatihan yang paling optimal. Meskipun akurasi pada data uji identik (99.34%) di semua skenario, skor CV yang lebih tinggi memberikan kepercayaan lebih bahwa model ini lebih robust terhadap variasi data.

Konsistensi parameter terbaik (`C=100`, `gamma=0.1`) di seluruh skenario menunjukkan bahwa pola data yang terbentuk cukup kuat dan tidak sensitif terhadap perubahan jumlah data latih dalam rentang 60-80%. Nilai `C=100` menunjukkan bahwa model memberikan penalti yang cukup besar terhadap kesalahan klasifikasi (margin yang ketat), sementara `gamma=0.1` menunjukkan jangkauan pengaruh data latih yang moderat, cocok untuk menangkap batas keputusan non-linear tanpa *overfitting* berlebihan.

### Efektivitas Penanganan *Class Imbalance*

Tantangan utama dalam dataset ini adalah ketidakseimbangan kelas (1.236 Layak vs 270 Tidak Layak). Penggunaan teknik pembobotan kelas (class weighting) terbukti sangat efektif. Hal ini tercermin dari nilai weighted avg untuk Precision, Recall, dan F1-Score yang semuanya berada di atas 99%.

Jika model bias terhadap kelas mayoritas, biasanya nilai Recall untuk kelas minoritas akan rendah. Namun, hasil *classification report* menunjukkan bahwa model mampu mengenali kedua kelas dengan sangat baik, meminimalkan risiko *False Negative* (air tidak layak diprediksi layak) yang berbahaya bagi kesehatan.

### Jawaban atas Rumusan Masalah

Mengacu pada rumusan masalah penelitian, yaitu "Seberapa efektif model SVM dapat diterapkan dalam mengklasifikasi status kelayakan air minum?", hasil penelitian ini menunjukkan bahwa SVM dengan kernel RBF **sangat efektif**. Dengan **akurasi** mencapai **99.34%** dan **F1-Score 99.34%**, model ini membuktikan keandalannya dalam mengklasifikasi kualitas air berdasarkan parameter fisik (Suhu, TDS, pH, Salinitas, Turbiditas) yang diambil menggunakan sensor murah.

## Implementasi Model Akhir

Berdasarkan analisis di atas, model final dibangun menggunakan parameter terbaik (`kernel='rbf'`, `C=100`, `gamma=0.1`) dan dilatih ulang menggunakan keseluruhan dataset (100% data) untuk memaksimalkan informasi yang dipelajari. Model yang telah dilatih beserta pipeline preprocessing disimpan ke dalam file format pickle (`water\_quality\_model.pkl`) agar dapat diimplementasikan lebih lanjut pada sistem embedded atau aplikasi pemantauan.

# BAB V KESIMPULAN

## Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut:

* Penggunaan teknik penyeimbangan bobot kelas (class weighting) berhasil mengatasi masalah ketidakseimbangan data (imbalanced dataset), sehingga model dapat mendeteksi kelas minoritas ("Tidak Layak") dengan akurat tanpa bias ke kelas mayoritas.
* Skenario pembagian data 70% training : 30% testing memberikan hasil validasi silang (Cross-Validation) terbaik sebesar 0.9972, menjadikannya skema yang paling optimal untuk dataset ini.
* Melalui skema optimalisasi parameter menggunakan GridSearchCV, ditemukan bahwa konfigurasi parameter terbaik adalah C = 100 dan gamma = 0.1. Konfigurasi ini konsisten memberikan performa terbaik pada berbagai skenario pembagian data.
* Penerapan algoritma Support Vector Machine (SVM) dengan kernel RBF terbukti sangat efektif untuk mengklasifikasi kelayakan air minum. Model mencapai tingkat akurasi pengujian sebesar 99.34% dengan nilai Presisi, Recall, dan F1-Score rata-rata tertimbang juga sebesar 99.34%.

## Saran

Untuk pengembangan penelitian selanjutnya dan implementasi sistem yang lebih baik, disarankan beberapa hal berikut:

* **Penambahan Parameter:** Meskipun parameter fisik sudah memberikan akurasi tinggi, penambahan parameter biologi (seperti bakteri E. Coli) atau kimiawi lainnya akan meningkatkan validitas standar kesehatan air secara medis.
* **Perluasan Dataset**: Menambah variasi lokasi pengambilan data di luar dua desa sampel saat ini akan membantu model menjadi lebih general (generalizable) terhadap karakteristik air di wilayah geografis yang berbeda.

# DAFTAR PUSTAKA

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M., & El-Shafie, A. (2019). Machine learning methods for better water quality prediction. Water, 11(6), 1231. https://doi.org/10.3390/w11061231

Atmaja, D. M. (2019). Analisis Kualitas Air Sumur Di Desa Candikuning Kecamatan Baturiti. *Media Komunikasi Geografi*, *19*(2), 147. https://doi.org/10.23887/mkg.v19i2.14644

Ali, M. M., Islam, M. S., Islam, A. R. M. T., Al-Ahmadi, S., Al-Ghamdi, A. A., & Youru, X. (2024). Advanced machine learning models for robust water quality index and water quality classification prediction. Journal of Hydroinformatics, 27(2), 299-322. https://doi.org/10.2166/hydro.2024.137

Arshad, M., Ali, I., Ahmad, M., Yasin, G., Hussain, M., Ali, H., & Khan, M. (2021). Optimization of operating parameters for bio-oil from rice husk ash using artificial neural network and support vector machine. Fuel, 305, 121516.

Baskar G. (2025). Machine Learning Approaches For Water Quality Classification Of Indian Water Bodies: A Comparative Analysis Of Random Forest, Svm, And Gradient Boosting. Sgs - Engineering & Sciences, 1(2).

Dzakwan, M. D. A. R., & Subektiningsih. (2023). Klasifikasi Tingkat Risiko Kesehatan Ibu Hamil Menggunakan Algoritma Support Vector Machine. *Indonesian Journal of Computer Science*, *12*(5). https://doi.org/10.33022/ijcs.v12i5.3372

G., L., & Villarica, M. (2019). Water quality monitoring and prediction in water reservoirs using data mining. IOP Conference Series: Materials Science and Engineering, 482, 012028.

Hassan, M. M., Hassan, M. M., Chowdhury, M. A., & Alam, M. M. (2021). Predicting and classifying water quality using machine learning algorithms. Water, 13(8), 4259. https://doi.org/10.3390/w13084259

Kencanawati, M., & Mustakim. (2017). Analisis Pengolahan Air Bersih Pada WTP PDAM Prapatan Kota Balikpapan. *Jurnal TRANSUKMA*, *02*(02), 2502–1028. http://transukma.uniba-bpn.ac.id/index.php/transukma/article/view/51/32

Kospa, H. S. D., & Rahmadi, R. (2019). Pengaruh Perilaku Masyarakat Terhadap Kualitas Air di Sungai Sekanak Kota Palembang. *Jurnal Ilmu Lingkungan*, *17*(2), 212. https://doi.org/10.14710/jil.17.2.212-221

Lee, A., & Holdijk, A. (2025). Predicting water potability using machine learning. Emerging Investigators.

Liang, J. Z. (2004). SVM multi-classifier and web document classification. *Proceedings of 2004 International Conference on Machine learning and Cybernetics*, *3*, 1347–1351. https://doi.org/10.1109/icmlc.2004.1381982

Maulina Putri, S. S., Arhami, M., & Hendrawaty, H. (2023). Penerapan Metode SVM pada Klasifikasi Kualitas Air. *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, *3*(2), 93. https://doi.org/10.30811/jaise.v3i2.4630

Nursaini, D., & Harahap, A. (2022). Kualitas Air Sungai. *BIOEDUSAINS:Jurnal Pendidikan Biologi Dan Sains*, *5*(1), 312–321. https://doi.org/10.31539/bioedusains.v5i1.3519

Patel, P., Singh, S., & Kumar, S. (2025). Comparative analysis of machine learning models for water quality classification. Journal of Mines, Metals and Fuels, 73(6), 1765-1772.x`

Permenkes RI. (2010). Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/Menkes/Per/IV/2010 Tentang Persyaratan Kualitas Air Minum. In *Peraturan Mentri Kesehatan Republik Indonesia* (p. MENKES).

Rahman, A., Astuti, L. P., Warsa, A., & Sentosa, A. A. (2021). Prediction of Water Turbidity Using Satellite Imagery Sentinel-2A in Jatiluhur Reservoir, West Java. *Jurnal Sumber Daya Air*, *17*(2), 59–68.

Ramon, E., Nazir, A., Novriyanto, N., Yusra, Y., & Oktavia, L. (2022). Klasifikasi Status Gizi Bayi Posyandu Kecamatan Bangun Purba Menggunakan Algoritma Support Vector Machine (SVM). *Jurnal Sistem Informasi Dan Informatika (Simika)*, *5*(2), 143–150. <https://doi.org/10.47080/simika.v5i2.2185>

Saleh, A. I., Al-Anzi, F. S., & Al-Anzi, A. F. (2024). Machine learning-based water quality prediction with grid search hyperparameter optimization. Journal of King Saud University - Computer and Information Sciences, 36(1), 101894.

Sharma, T., & Sharma, R. (2024). Machine learning framework for predicting water quality classification. Water Practice & Technology, 19(11), 4499-4513.

Singh, K. P., Basant, N., Malik, A., & Jain, G. (2011). Support vector machines in water quality management. Analytica Chimica Acta, 703(2), 152-162.

Sudin, A., Salmin, M., Fhadli, M., & ... (2023). Klasifikasi Kelayakan Air Minum Bagi Tubuh Manusia Menggunakan Metode Support Vektor Machine Dengan Backward Elimination. *Jurnal Jaringan Dan …*, *3*(1), 87–95. https://doi.org/00.0000/jati

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