

LAPORAN AKHIR
STUDI INDEPENDEN BERSERTIFIKAT
Foundations of AI and Life Skills for Gen-Z
Di Orbit Future Academy

Diajukan untuk memenuhi persyaratan kelulusan
Program MSIB MBKM

oleh :
Sukma Nur Savitri / A410190137



Program Studi Pendidikan Matematika
Fakultas Keguruan dan Ilmu Pendidikan
Universitas Muhammadiyah Surakarta
2022

Lembar Pengesahan

Program Studi Pendidikan Matematika

Universitas Muhammadiyah Surakarta

**Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma
*Machine Learning***

Di Orbit Future Academy

oleh :

Sukma Nur Savitri / A410190137

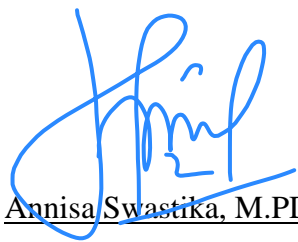
disetujui dan disahkan sebagai

Laporan Magang atau Studi Independen Bersertifikat Kampus Merdeka

Surakarta, 16 Juni 2022

Pembimbing Magang atau Studi Independen

Program Studi Pendidikan Matematika & Universitas Muhammadiyah Surakarta



Annisa Swastika, M.PD.

NIDN: 0613098601

Lembar Pengesahan

**Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma
*Machine Learning***

Di Orbit Future Academy

oleh :

Sukma Nur Savitri / A410190137

disetujui dan disahkan sebagai

Laporan Magang atau Studi Independen Bersertifikat Kampus Merdeka

Bandung, 16 Juni 2022

AI Coach

Angel Metanosa Afinda, S.Kom.

NIP: 2201043

Abstraksi

Foundation of AI and Life-skills for Gen-Z adalah program MSIB (Magang dan Studi Independen Bersertifikat) Kampus Merdeka di Perusahaan Orbit Future Academy. *Foundation of AI and Life Skills for Gen-Z* merupakan program pelatihan *Artificial Intelligence* secara *online* untuk mahasiswa yang bertujuan bukan hanya untuk memperkenalkan teknologi AI ke mahasiswa, tetapi juga untuk memungkinkan mereka bisa mengangkat perangkat AI, sehingga bisa membuat sesuatu produk yang menciptakan dampak sosial. OFA berfokus pada komponen utama AI seperti *Data Science, Natural Language Processing and Computer Vision*. Materi pembelajaran meliputi Logika dan Konsep Teknologi AI, Metode Penelitian AI, Siklus Projek AI, Pemrograman *Python*, Etika Profesi dan Keterampilan Perusahaan, Keterampilan Kewirausahaan, dan Proyek Akhir. Selain keterampilan AI, pada Program *Foundations of AI and Life Skills for Gen-Z* akan mengajarkan keterampilan hidup yang bisa dipergunakan untuk mencari pekerjaan atau menciptakan pekerjaan. Tujuan dari program ini adalah agar dapat meningkatkan keterampilan dan wawasan teknologi generasi muda Indonesia serta agar mampu bersaing dengan Negara lain di bidang industri dan teknologi. Laporan akhir ini membahas Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma *Machine Learning* untuk menambah wawasan terkait kualitas air minum yang baik.

Kata Kunci: *Artificial Intelligence, Algoritma, Machine Learning, Klasifikasi, Kualitas Air Minum*

Kata Pengantar

Segala puji syukur kehadiran Allah Swt., yang telah melimpahkan rahmat dan hidayah-Nya kepada kita semua, sehingga penyusun dapat membuat laporan Proyek Akhir untuk memenuhi tugas akhir dari Orbit Future Academy sebagai syarat kelulusan program studi independen bersertifikat kampus merdeka. Penyusun berusaha dengan semaksimal mungkin demi kesempurnaan penyusunan laporan ini baik dari hasil kegiatan belajar di Orbit Future Academy (OFA) dan bekerja secara Tim dalam pengerjaan Proyek Akhir. Saran dan kritik yang sifatnya membangun begitu diharapkan oleh penyusun demi kesempurnaan dalam penulisan laporan berikutnya.

Kesempatan ini, penyusun mengucapkan banyak terima kasih kepada semua pihak yang telah membantu dan menyelesaikan dalam penyusunan Laporan Proyek Akhir Program Studi Independen Bersertifikat di antaranya:

1. Kepada Rekan/Tim yang bekerja sama dalam pengerjaan Laporan Proyek Akhir.
2. Kepada Coach Angel Metanosa Afinda, S.Kom., selaku Pembimbing di kampus Mitra.
3. Kepada Annisa Swastika, M.PD., selaku Dosen Pembimbing di Program Studi Pendidikan Matematika Universitas Muhammadiyah Surakarta.

Akhir kata, penyusun berharap laporan ini dapat bermanfaat bagi pembaca serta dapat dijadikan motivasi dalam mengembangkan Proyek Akhir kedepannya. Penyusun ucapkan terima kasih kepada semua pihak yang telah membantu, semoga Allah Swt. membalas semua kebaikan kalian. Aamiin.

Kendal, 16 Juni 2022



Penyusun

Daftar Isi

Bab I Pendahuluan	1
I.1 Latar belakang	1
I.2 Lingkup	2
I.3 Tujuan	2
Bab II Orbit Future Academy	1
II.1 Struktur Organisasi	1
II.2 Lingkup Pekerjaan	2
II.3 Deskripsi Pekerjaan	3
II.4 Jadwal Kerja	4
Bab III Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning	1
III.1 Latar Belakang Proyek Akhir	1
III.2 Proses Pelaksanaan Proyek Akhir	3
III.3 Hasil Proyek Akhir	10
Bab IV Penutup	vii
IV.1 Kesimpulan	vii
IV.2 Saran	vii
Bab V Referensi	viii
Bab VI Lampiran A. TOR	1
Bab VII Lampiran B. Log Activity	3
Bab VIII Lampiran C. Jurnal	1

Daftar Tabel

Tabel 2.1 Agenda Kelas	4
Tabel 3.1 Hasil Klasifikasi	10

Daftar Gambar

Gambar 2.1 Logo Orbit Future Academy	1
Gambar 2.2 Struktur Organisasi OFA	2
Gambar 3.1 Rumus dan Grafik <i>Logistic Regression</i>	7
Gambar 3.2 Alur Proses Algoritma SVM	7
Gambar 3.3 Alur Random Forest	8
Gambar 3.4 Diagram Kualitas Air Minum berdasarkan Database	10
Gambar 3.5 <i>Confusion Matrix Random Forest Classifier</i>	11

Bab I Pendahuluan

I.1 Latar belakang

Program Magang dan Studi Independen Bersertifikat (MSIB) adalah bagian dari program Kampus Merdeka yang bertujuan untuk memberikan kesempatan kepada mahasiswa untuk belajar dan mengembangkan diri melalui aktivitas di luar kelas perkuliahan, namun tetap diakui sebagai bagian dari perkuliahan. Program studi independen bertujuan untuk mempelajari kompetensi yang spesifik, praktis, dan dibutuhkan di masa depan. Selain itu, program studi independen sendiri memudahkan peserta dalam mengembangkan suatu hal dikarenakan akan berinteraksi langsung dengan pakar yang memahami penerapannya, hal ini juga akan membantu siswa dalam memahami hal baru di luar kampus dengan menerapkan kompetensi dengan berbasis proyek riil.

Program studi independen melakukan kerjasama dengan mitra yang sesuai dengan bidangnya seperti Perusahaan Orbit Future Academy yang berfokus pada pengetahuan mengenai kecerdasan buatan. Perusahaan Orbit Future Academy didirikan dengan tujuan untuk meningkatkan kualitas hidup melalui inovasi, pendidikan, dan pelatihan keterampilan. Program studi independen *Foundation of AI and Life Skills for Gen-Z* di Perusahaan Orbit Future Academy berlangsung selama 5 bulan mulai 21 Februari sampai 22 Juli 2022, berfokus pada komponen utama *Artificial Intelligence* seperti *Data Science, Natural Language Processing and Computer Vision*. Materi pembelajaran meliputi Logika dan Konsep Teknologi AI, Metode Penelitian AI, Siklus Proyek AI, Pemrograman Python, Etika Profesi & Keterampilan Perusahaan, Keterampilan Kewirausahaan, dan Proyek Akhir. Selain keterampilan AI, pada Program Foundations of AI and Life Skills for Gen-Z akan mengajarkan keterampilan hidup yang bisa dipergunakan untuk mencari kerja atau menciptakan kerja.

Selama program *Foundation of AI and Life Skills for Gen-Z*, kegiatan pembelajaran dilaksanakan secara *online* atau daring dengan software Microsoft Teams dan Zoom Meeting. Jenis kegiatan berupa penjelasan materi oleh *homeroom*

coach atau *domain coach*, sesi diskusi, *live coding*, serta tugas untuk latihan. Hasil program *Foundation of AI and Life Skills for Gen-Z* di Perusahaan Orbit Future Academy menekankan peserta dapat membuat suatu proyek berbasis dengan kecerdasan buatan sebagai output dari pembelajaran yang sudah dilakukan. Sebelum pembuatan proyek para peserta akan diberi pembelajaran mengenai AI dan domain AI yang dapat digunakan dalam mengembangkan proyek sebagai luaran hasil akhir.

I.2 Lingkup

Mengacu pada Lampiran A. TOR, mentee berkewajiban untuk mengikuti kegiatan berikut dalam portofolio-nya:

1. Kegiatan Tim Pemograman Python merupakan kegiatan yang dilakukan secara berkelompok oleh masing-masing anggota kelompok berjumlah 4-5 anggota. Setiap kelompok membuat model pemograman sesuai dengan yang telah dipilih dan dikonsultasikan bersama dengan AI Coach kelas.
2. Program Orbit Future Academy memberikan pelatihan berupa challenge kepada meteenya sebagai umpan balik terhadap mahasiswa agar mampu kembali menjelaskan dan mengerjakan apa yang sudah mentee pelajari dari AI Coach.

I.3 Tujuan

Berikut tujuan serta hasil yang diperoleh selama mengikuti MSIB di OFA:

1. Memiliki wawasan tentang AI dan perkembangannya.
2. Mampu merancang dan mengimplementasikan AI Project Cycle.
3. Mampu menggunakan bahasa pemrograman Python untuk mengembangkan aplikasi AI.
4. Mampu mengembangkan salah satu dari 3 domain AI (DS, NLP, dan CV) hingga tahap *deployment*.
5. Mampu menggunakan *soft skills* dan *hard skills* dalam dunia industri dan lingkungan perusahaan.

6. Mampu mengaplikasikan kiat-kiat yang dibutuhkan seorang wirausahawan yang bergerak di bidang *start-up* dalam mentransformasikan ide ke dalam bentuk produk/jasa sehingga dapat menciptakan peluang bisnis yang terus berinovasi, berevolusi, dan berkelanjutan.

Bab II Orbit Future Academy

II.1 Struktur Organisasi



Gambar 2.1 Logo Orbit Future Academy

Orbit Future Academy (OFA) didirikan pada tahun 2016 dengan tujuan untuk meningkatkan kualitas hidup melalui inovasi, edukasi, dan pelatihan keterampilan. Label atau *brand* Orbit merupakan kelanjutan dari warisan mendiang Prof. Dr. Ing. B. J. Habibie (presiden Republik Indonesia ke-3) dan istrinya, Dr. Hasri Ainun Habibie. Mereka berdua telah menjadi penggerak dalam mendukung perkembangan inovasi dan teknologi pendidikan di Indonesia. OFA mengkurasi dan melokalkan program/kursus internasional untuk *upskilling* atau *reskilling* pemuda dan tenaga kerja menuju pekerjaan masa depan. Hal ini sesuai dengan slogan OFA, yakni “*Skills-for-Future-Jobs*”.

Visi:

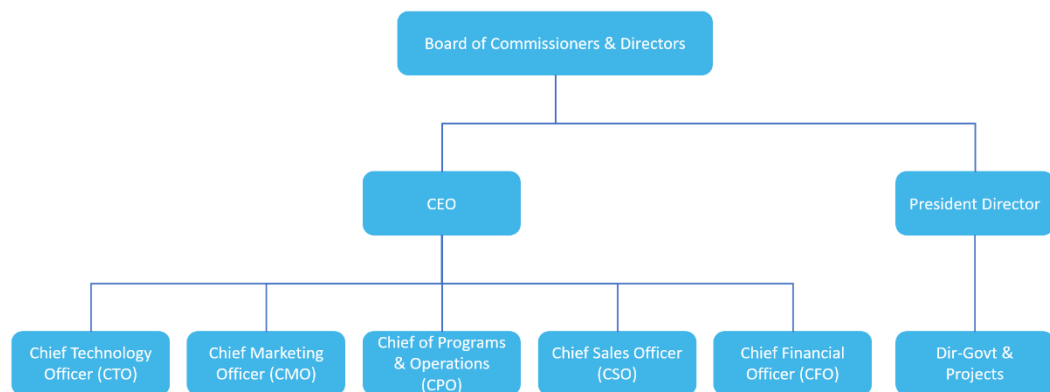
Memberikan pembelajaran berbasis keterampilan transformatif terbaik untuk para pencari kerja & pencipta lapangan kerja.

Misi:

1. Membangun jaringan Orbit Transformation Center (OTC) secara nasional untuk menyampaikan kurikulum keterampilan masa depan berbasis sertifikasi melalui Platform Konten Digital.

2. Secara proaktif bekerja dengan pemerintah & organisasi dengan mengubah tenaga kerja mereka agar sesuai dengan perubahan pekerjaan yang terjadi karena Industri 4.0.
3. Melatih pemuda dengan keterampilan kewirausahaan & mencocokkan mereka dengan peluang masa depan yang muncul di berbagai industri.
4. Menghubungkan jaringan inkubator dan akselerator yang dikurasi ke industri, investor, dan ekosistem start-up global.

Struktur organisasi OFA dapat dilihat pada Gambar 2.2.



Gambar 2.2 Struktur Organisasi OFA

II.2 Lingkup Pekerjaan

Seorang fasilitator akan mendampingi kurang lebih 40 peserta MSIB (student) dalam satu kelas. Terdapat dua jenis fasilitator, yakni:

a. Homeroom Coach

Homeroom coach bertugas menyampaikan materi tentang dasar-dasar AI, memberikan penilaian pada student, dan mengarahkan *student* saat pengerjaan Proyek Akhir (PA).

b. Domain Coach

Domain coach bertugas menyampaikan materi tentang domain AI atau life skills dan memberikan penilaian pada student.

Lingkup pekerjaan student adalah mengikuti kelas bersama homeroom atau domain coach, sesuai agenda kelas, hingga program selesai.

II.3 Deskripsi Pekerjaan

Berikut adalah deskripsi pekerjaan student sebelum pengerjaan PA:

- a. Mengikuti pre-test.
- b. Mengikuti kelas sesi pagi pada pukul 08.00 hingga 11.30 WIB.
- c. Mengikuti kelas sesi siang pada pukul 13.00 hingga 16.30 WIB.
- d. Mengulang materi yang telah disampaikan di kelas sesi pagi dan siang, setelah kelas sesi siang, selama 1 jam (*self-study*).
- e. Mengerjakan latihan individu atau kelompok yang diberikan oleh homeroom atau domain coach saat kelas berlangsung.
- f. Mengerjakan tugas yang diberikan homeroom atau domain coach hingga batas waktu tertentu.
- g. Mengerjakan *mini project* yang diberikan homeroom atau domain coach hingga batas waktu tertentu
- h. Mengikuti post-test.

Student memiliki peran membuat codingan dan menyusun hasil pembahasan jurnal selama pengerjaan PA, dengan deskripsi pekerjaan sebagai berikut:

- a. Membuat dan merevisi codingan proyek akhir dari awal sampai akhir dengan bantuan anggota lainnya.
- b. Menyusun jurnal bagian hasil dan pembahasan.
- c. Menyusun laporan akhir.

II.4 Jadwal Kerja

Program ini berlangsung setiap hari kerja (Senin sampai dengan Jumat) selama 8 jam per harinya, dengan rincian sebagai berikut:

Tabel 2.1 Agenda Kelas

Pukul (WIB)	Durasi (jam)	Aktivitas
08.00 s.d. 11.30	3.5	Kelas Sesi Pagi
13.00 s.d. 16.30	3.5	Kelas Sesi Siang
16.30 s.d. 17.30	1	<i>Self-Study</i>

Program ini berlangsung dari bulan Februari 2022 sampai dengan bulan Juli 2022.

Bab III Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma *Machine Learning*

III.1 Latar Belakang Proyek Akhir

Air bersih merupakan kebutuhan dasar bagi manusia, sehingga ketersediaannya amatlah penting. Air dimanfaatkan dalam kesehariannya tidak hanya terbatas untuk keperluan rumah tangga, tetapi juga untuk fasilitas umum, sosial dan ekonomi. Kebutuhan air bersih terus meningkat seiring dengan perkembangan populasi manusia. Kebutuhan dan permintaan air bersih akan terus meningkat dengan adanya pertumbuhan penduduk, terjadi pergerakan dinamik dalam masyarakat baik dalam segi kepadatan, sosial maupun ekonomi. Hal tersebut membuat manusia terikat dengan fungsi air. Akan tetapi tidak semua zat-zat mineral yang terkandung dalam air dapat dikonsumsi dengan layak.

Air rentan terkena oleh bakteri yang sangat berbahaya bagi sistem pencernaan manusia. Hal ini air terbagi menjadi dua yaitu air yang layak konsumsi dan tidak layak konsumsi. Berdasarkan Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/MENKES/PER/IV/2010 tentang Persyaratan Kualitas Air Minum terdapat pengertian mengenai Air Minum yaitu air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum [1]. Identifikasi dini terhadap produk air dari sumber air baku sangat diperlukan untuk mengetahui faktor-faktor yang mempengaruhi kualitas air yang layak konsumsi bagi masyarakat.

Krisis air bersih sedang melanda berbagai negara di dunia dengan hanya 1% jumlah air bersih yang dapat dikonsumsi oleh manusia. Jumlah yang sangat kecil untuk jumlah air bersih yang baik, menyebabkan air bersih susah diakses oleh penduduk. Sebanyak 663 juta penduduk yang bersumber dari data WHO menunjukkan bahwa mereka susah untuk mengakses air bersih [4]. Berdasarkan data UNESCO, pada tahun 2025 diprediksi bahwa dua pertiga dari jumlah penduduk dunia akan tinggal di daerah yang kekurangan air bersih [4]. Posisi *World Water Asssment Programme* (WWAP) dibawah UNESCO sudah melakukan prediksi akan kondisi air bersih untuk beberapa tahun kedepan. Pada kondisi

tertentu, sebagai gambaran kebutuhan akan air dalam kehidupan sehari-hari tidak kurang dari 85% air bersih berubah menjadi air limbah. Setiap orang bisa menggunakan hingga 100 liter air perhari untuk memenuhi kebutuhan hidupnya [4].

Semakin hari juga tingkat kualitas dari air tersebut semakin menurun. Jika hal itu terus terjadi kondisi seperti ini sangat fatal dan dapat mengakibatkan gangguan pada pencernaan manusia dan seluruh makhluk hidup yang membutuhkan air tersebut. Upaya dilakukan dalam menjaga kualitas air tersebut dengan melakukan pengolahan dengan cara pengawasan di daerah air atau lingkungan sumber daya air. Hal ini berguna agar sumber air tetap terjaga kelestariannya dan dapat menghasilkan kualitas air dengan standart air bersih yang layak dikonsumsi oleh manusia. Kualitas air bersih yang dimaksud yaitu melakukan pengukuran kondisi terhadap air berdasarkan karakteristik pada fisik air, karakteristik kimiawi dan biologisnya. Pengukuran tersebut untuk mengukur kualitas air tersebut memenuhi syarat yang layak dikonsumsi atau tidak dapat mengetahuinya melalui zat-zat dan mineral yang terkandung di dalam air minum. *Machine learning* lebih memudahkan untuk proses implementasi. *Machine Learning* merupakan salah satu cabang dari ilmu Kecerdasan Buatan, khususnya yang mempelajari tentang bagaimana komputer mampu belajar dari data untuk meningkatkan kecerdasannya.

Menurut Prismahardi Aji Riyantoko, Tresna Maulana Fahrudin, dan Kartika Maulida Hindrayani dalam penelitiannya yang berjudul “Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan menggunakan *Lucifer Machine Learning*” menyimpulkan bahwa metode *LuciferML* merupakan *semi-supervised machine learning* dimana gabungan dari *supervised* dan *unsupervised machine learning*. Tujuan pengembangan *library* tersebut adalah untuk mempermudah analisis model klasifikasi dengan akurat berdasarkan jenis-jenis data. Hasil yang diperoleh pada penelitian ini dapat dilihat dari model terbaik dengan tingkat akurasi tertinggi yaitu *Random Forest Classifier* sebesar 72,81%. Persamaan dengan penelitian ini yaitu menerapkan algoritma *machine learning* dengan perbedaannya terdapat pada hasil akurasi yang diperoleh algoritma *machine learning* yang digunakan. Bagaimana penggambaran dari kualitas air?

Penggambaran kualitas air biasanya digambarkan dalam bentuk parameter dan variabel. Beberapa parameter bermacam-macam digunakan sebagai dasar untuk menentukan model seperti apa *paper* ini disusun.

Parameter-parameter tersebut dapat digunakan untuk membuat permodelan klasifikasi dan memprediksi nilai parameter yang lainnya, yang termuat dalam *paper*. Dalam proses memprediksi data dapat digunakan beberapa metode klasifikasi baik secara manual maupun komputasional dengan memanfaatkan *machine learning*. Beberapa algoritma pemodelan yang digunakan pada penelitian ini adalah *Logistic regression*, *SVM*, *Random Forest Classifier*, *KNN*, dan *XGB Classifier*. Penelitian ini, akan menggunakan *semi-supervised machine learning* dengan menggunakan *library python* yang digunakan untuk mengolah data tabular. Metode tersebut digunakan untuk melakukan analisis dan membantu untuk melakukan pre-prosesing data dalam melakukan prediksi dan klasifikasi pada kualitas air minum.

III.2 Proses Pelaksanaan Proyek Akhir

Berikut proses pelaksanaan proyek akhir Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma *Machine Learning*.

2.1 Dataset

Proses pelaksanaan proyek akhir ini menggunakan data yang bersumber dari situs *website* yaitu Kaggle yang bernama *water potability* yang berformat csv yang di selesaikan dengan menggunakan algoritma *Machine Learning*. Data tersebut memiliki sepuluh parameter diantaranya sebagai berikut.

a. pH

pH (*Power of Hydrogen*) adalah skala yang digunakan untuk menyatakan tingkat keasaman atau kebasaan yang dimiliki oleh suatu larutan. Skala dari pH terdiri dari angka 1 hingga 14. WHO telah merekomendasikan batas maksimum pH yang diizinkan dari 6,5 hingga 8,5. Standar kisaran dari WHO tentang investigasi saat ini adalah 6,52-6,83.

b. Hardness

Hardness adalah kandungan mineral-mineral tertentu di dalam air. Umumnya ion kalsium (Ca) dan magnesium (Mg) dalam bentuk garam karbonat.

c. Solids (Total Padatan Terlarut – TDS)

Istilah untuk menandakan jumlah padatan terlarut atau konsentrasi jumlah ion kation (bermuatan positif) dan anion (bermuatan negatif) di dalam air. Air memiliki kemampuan untuk melarutkan berbagai jenis anorganik dan beberapa mineral atau garam organik seperti kalium, kalsium, natrium, bikarbonat, klorida, magnesium, sulfat, dan lain sebagainya. Mineral ini menghasilkan rasa yang tidak diinginkan dan warna yang encer dalam penampilan air. Solids adalah parameter penting untuk penggunaan air. Air dengan nilai TDS yang tinggi menunjukkan bahwa air tersebut sangat termineralisasi. Batas yang diinginkan untuk TDS adalah 500 mg/l dan batas maksimum adalah 1000 mg/l yang ditentukan untuk tujuan minum.

d. Chloramines

Sebuah kompleks kimia yang terdiri dari klorin dan amonia. Klorin dan kloramin merupakan disinfektan utama yang digunakan dalam sistem air publik. Kloramin paling sering terbentuk ketika amonia ditambahkan ke klorin untuk mengolah air minum. Kadar klorin hingga 4 miligram per liter (mg/L atau 4 bagian per juta (ppm)) dianggap aman dalam air minum.

e. Sulfate

Sulfat merupakan kandungan foaming agent (mampu menimbulkan busa) dan biasa dipakai pada produk seperti pembersih wajah, sampo, dan pasta gigi. Sulfat adalah zat alami yang ditemukan di mineral, tanah, dan batuan. Penggunaan komersial utama sulfat adalah dalam industri kimia. Konsentrasi sulfat dalam air laut adalah sekitar 2.700 miligram per liter (mg/L). Ini berkisar antara 3 sampai 30 mg/L di sebagian besar

persediaan air tawar, meskipun konsentrasi yang jauh lebih tinggi (1000 mg/L) ditemukan di beberapa lokasi geografis.

f. Conductivity

Konduktivitas adalah ukuran kemudahan di mana muatan listrik atau panas dapat melewati suatu bahan. Air murni bukan penghantar arus listrik yang baik, tetapi isolator yang baik. Peningkatan konsentrasi ion meningkatkan konduktivitas listrik air. Umumnya, jumlah padatan terlarut dalam air menentukan konduktivitas listrik. Konduktivitas listrik (EC) sebenarnya mengukur proses ionik suatu larutan yang memungkinkannya mentransmisikan arus. Standar dari WHO, nilai EC tidak boleh melebihi 400 S/cm.

g. Organic Carbon

Karbon memiliki TOC yang digunakan untuk mengetahui jumlah total carbon dalam air murni. Total Organic Carbon (TOC) di perairan sumber berasal dari bahan organik alami (NOM) yang membusuk serta sumber sintetis. TOC adalah ukuran jumlah total karbon dalam senyawa organik dalam air murni dengan < 2 mg/L sebagai TOC dalam air olahan/minum, dan < 4 mg/Lit pada air sumber yang digunakan untuk pengolahan.

h. Trihalomethanes

Trihalomethanes (THMs) adalah hasil reaksi antara klorin yang digunakan untuk mendisinfeksi air keran dan bahan organik alami di dalam air. Pada tingkat tinggi, THMs telah dikaitkan dengan efek kesehatan negatif seperti kanker dan hasil reproduksi yang merugikan. Kadar dari THM hingga 80 ppm dianggap aman dalam air minum.

i. Turbidity

Turbidity (kekeruhan) adalah ukuran kejernihan relatif suatu cairan. Kekeruhan air tergantung pada jumlah zat padat yang ada dalam keadaan tersuspensi. Nilai rata-rata kekeruhan yang diperoleh untuk Kampus Wondo Genet (0,98 NTU) lebih rendah dari nilai rekomendasi WHO sebesar 5,00 NTU.

j. *Potability*

Parameter yang menunjukkan jika air yang layak konsumsi berkode 1 dan tidak dapat konsumsi berkode 0.

2.2 *Missing Value*

Missing Value terjadi ketika ada dataset yang bernilai NaN. Dataset yang kami pilih terdapat *missing value* pada kolom pH, Sulfat, dan trihalomethanes. Banyak cara yang bisa dilakukan untuk menghilangkan terjadinya *missing value*. Substitusi nilai rata-rata merupakan salah satu cara untuk menghilangkan terjadinya *missing value*. Cara tersebut kami pilih untuk menangani terjadinya *missing value*. Setelah mensubstitusikan nilai rata-rata dataset yang kami miliki sudah tidak terjadi *missing value*.

2.3 *Clasification*

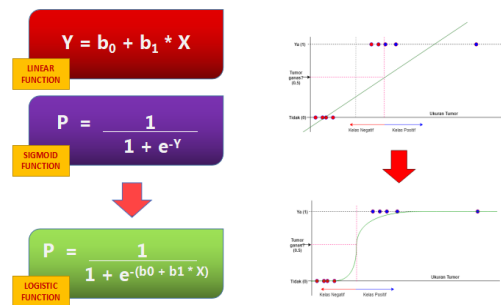
Classification dalam *data science* berarti proses memprediksi kelas atau kategori data dengan memanfaatkan nilai yang ada pada data. Algoritma *machine learning* terbagi menjadi dua, yaitu *supervised* dan *unsupervised learning*. *Classification* termasuk dalam algoritma *supervised learning*. Algoritma yang digunakan dalam *classification* sangat beragam, seperti *logistic regression*, *random forest*, dan lain-lain. Proses *classification* pada dasarnya dilakukan agar analisis data menjadi lebih mudah dan tentunya memberikan hasil yang akurat. Agar bisa memberikan suatu informasi yang bermanfaat.

2.4 Algoritma *Machine Learning*

1. *Logistic Regression*

Logistic regression adalah jenis analisis statistik yang sering digunakan data analyst untuk pemodelan prediktif. Pendekatan analitik ini, variabel dependennya terbatas atau kategoris, berupa A atau B (regresi biner) atau berbagai opsi hingga A, B, C atau D (regresi multinomial). Jenis analisis statistik digunakan dalam

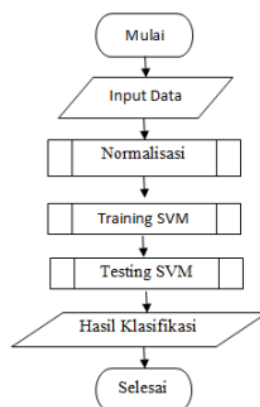
software statistik untuk memahami hubungan antara variabel dependen dan satu atau lebih variabel independen dengan memperkirakan probabilitas. Jenis analisis ini dapat membantu Anda memprediksi kemungkinan.



Gambar 3.1 Rumus dan Grafik *Logistic Regression*

2. *Support Vector Machine (SVM)*

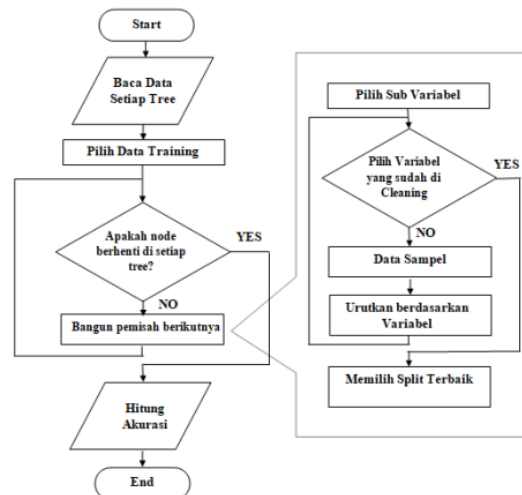
Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di *Annual Workshop on Computational Learning Theory* [2]. *Support Vector Machine (SVM)* adalah algoritma pembelajaran *machine* terawasi yang dapat digunakan untuk tantangan klasifikasi atau regresi. Namun, sebagian besar SVM digunakan dalam masalah klasifikasi. Algoritma SVM, kami memplot setiap item data sebagai titik dalam ruang n -dimensi (n adalah sejumlah fitur yang Anda miliki) dengan nilai setiap fitur menjadi nilai koordinat tertentu.



Gambar 3.2 Alur Proses Algoritma SVM

3. *Random Forest*

Random Forest adalah pengklasifikasi yang berisi sejumlah pohon keputusan pada berbagai subset dari dataset yang diberikan dan mengambil rata-rata untuk meningkatkan akurasi prediksi dari dataset itu. Metode random forest (RF) merupakan metode yang dapat meningkatkan hasil akurasi, karena dalam membangkitkan simpul anak untuk setiap node dilakukan secara acak [5]. Metode *random forest* digunakan untuk membangun pohon keputusan yang terdiri dari *root node*, *internal node*, dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan.



Gambar 3.3 Alur Algoritma Random Forest

4. *K-Nearest Neighbor (KNN)*

Berdasarkan teknik *Supervised Learning* *K-Nearest Neighbor* merupakan salah satu algoritma *Machine Learning* yang paling sederhana. Algoritma *K-Nearest Neighbor* pertama kali ditemukan pada awal tahun 1950-an didasarkan pada pembelajaran dengan analogi, yaitu dengan membandingkan *tuple* tes yang diberikan dengan *tuple* pelatihan yang serupa dengannya, *tuple* pelatihan dijelaskan oleh n atribut [6]. Algoritma ini mengasumsikan kesamaan antara kasus atau data baru dengan kasus yang telah tersedia dan memasukkan kasus baru ke dalam kategori yang paling mirip dengan

kategori yang tersedia. Algoritma K-NN menyimpan semua data yang tersedia dan mengklasifikasikan titik data baru berdasarkan kesamaannya. Artinya ketika data baru muncul maka dapat dengan mudah untuk diklasifikasikan ke dalam kategori *well suite* dengan menggunakan algoritma K-NN.

Salah satu langkah proses klasifikasi Algoritma *modified k-nearest neighbor* yaitu terkait perhitungan jarak *Euclidean*. Jarak ini berguna untuk mendefinisikan jarak antara dua titik yaitu titik pada data *training* (x) dan titik pada data *testing* (y) maka digunakan rumus *Euclidean*.

$$d(x_i, y_i) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2}$$

dengan d adalah jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi, dimana $x = x_1, x_2, \dots, x_i$ dan $y = y_1, y_2, \dots, y_i$ dan I merepresentasikan nilai atribut serta n merupakan dimensi atribut [6].

5. XGBoost

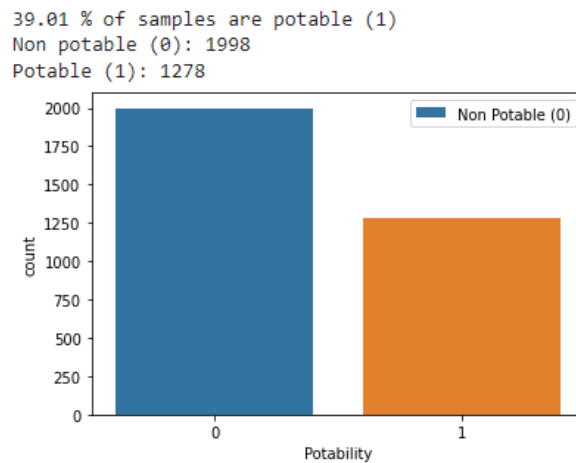
XGBoost adalah implementasi dari pohon keputusan yang didorong gradien yang dirancang untuk kecepatan dan kinerja. *XGBoost* adalah algoritma peningkatan gradien ekstrim. Hal ini berarti adalah algoritma pembelajaran mesin yang besar dengan banyak bagian. *XGBoost* bekerja dengan kumpulan data yang besar dan rumit. *XGBoost* adalah teknik pemodelan *ensemble*.

2.5 Confusion Matrix

Confusion matrix merupakan dataset yang memiliki dua kelas, kelas yang satu positif dan kelas lain sebagai negatif. Terdiri dari empat sel yaitu *true positif*, *false positif*, *true negative*, and *false negative* [3].

III.3 Hasil Proyek Akhir

Hasil dari *exploratory data analysis* total data sebanyak 3276 data dan 10 kolom parameter.



Gambar 3.4 Diagram Kualitas Air Minum berdasarkan Database

Berdasarkan hasil diagram di atas dari 3276 data bahwa terdapat 1998 *non potable* atau air tidak dikonsumsi dan 1278 *potable* atau air dapat dikonsumsi.

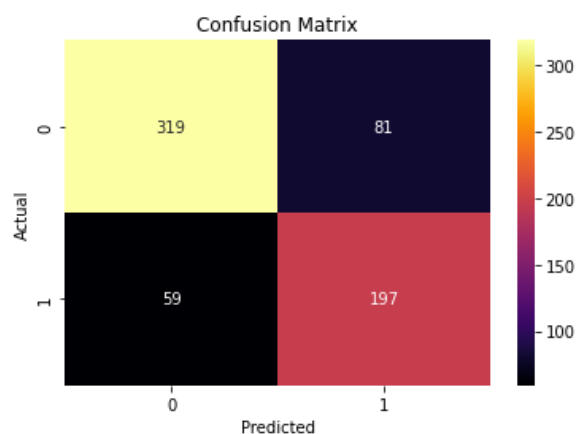
Proses klasifikasi kualitas air minum melewati proses modelling yang dilakukan dengan menggunakan lima algoritma *machine learning* menghasilkan nilai yang berbeda antara algoritma satu dengan yang lain. **Tabel 3.1** akan menunjukkan hasil dari proses modelling lima algoritma tersebut.

Tabel 3.1 Hasil Klasifikasi

<i>Classifier</i>	<i>Class</i>	<i>Precision</i>	<i>Recall</i>	<i>f1-score</i>	Akurasi
<i>Logistic Regression</i>	0	0.64	0.51	0.56	0.52
	1	0.42	0.55	0.47	
<i>SVM</i>	0	0.75	0.65	0.70	0.66
	1	0.55	0.66	0.60	
<i>XGB Classifier</i>	0	0.87	0.69	0.77	0.75
	1	0.64	0.84	0.73	
<i>Random Forest</i>	0	0.85	0.79	0.81	0.78
	1	0.70	0.78	0.74	
<i>KNN</i>	0	0.71	0.63	0.67	0.62
	1	0.51	0.61	0.56	

Berdasarkan hasil diatas bahwa pada *classifier logistic regression* memiliki *precision* kelas 0 adalah 64% dan kelas 1 adalah 42%, *recall* kelas 0 adalah 51% dan kelas 1 adalah 55%, *f1-score* kelas 0 adalah 56% dan kelas 1 adalah 47%, serta akurasi sebesar 52%. Pada *classifier SVM* memiliki *precision* kelas 0 adalah 75% dan kelas 1 adalah 55%, *recall* kelas 0 adalah 65% dan kelas 1 66%, *f1-score* kelas 0 adalah 70% dan kelas 1 adalah 60%, serta akurasi sebesar 66%. Pada *XGB classifier* memiliki *precision* kelas 0 adalah 87% dan kelas 1 adalah 64%, *recall* kelas 0 adalah 69% dan kelas 1 84%, *f1-score* kelas 0 adalah 77% dan kelas 1 adalah 73%, serta akurasi sebesar 74%. Pada *random forest classifier* memiliki *precision* kelas 0 adalah 85% dan kelas 1 adalah 70%, *recall* kelas 0 adalah 79% dan kelas 1 78%, *f1-score* kelas 0 adalah 81% dan kelas 1 adalah 74%, serta akurasi sebesar 78%. Pada *classifier KNN* memiliki *precision* kelas 0 adalah 71% dan kelas 1 adalah 51%, *recall* kelas 0 adalah 63% dan kelas 1 61%, *f1-score* kelas 0 adalah 67% dan kelas 1 adalah 56%, serta akurasi sebesar 62%.

Tingkat akurasi yang dihasilkan setiap model berbeda. Berdasarkan urutan model dengan tingkat akurasi terendah ke tertinggi dapat dilihat sebagai berikut *logistic regression*, *K-Nearest Neighbor*, *Support Vector Machine*, *XGB Classifier*, dan *Random Forest Classifier*. Oleh karena itu, dapat disimpulkan bahwa *Random Forest Classifier* memiliki akurasi tertinggi yaitu 78%.



Gambar 3.5 Confusion Matrix Random Forest Classifier

Berdasarkan Gambar 3.5 akan diketahui nilai pesisi, recall, dan akurasi pada *random forest classifier*. Nilai yang dihasilkan gambar diatas dapat diartika bahwa

nilai 319 yang berarti *True Positive* (TP), nilai 197 yang berarti *True Negative* (TN), nilai 81 berarti *False Positif* (FP), dan nilai 59 berarti *False Negative* (FN).

Presisi merupakan persentase kasus yang diprediksi positif yang ternyata benar. Presisi dapat dihitung dengan menggunakan rumus sebagai berikut.

$$\text{Presisi} = \frac{TP}{TP+FP} = \frac{319}{319+81} = \frac{319}{400} = 0.7975 = 79.75\%$$

Recall digunakan untuk mengukur pecahan kasus positif yang diidentifikasi dengan benar. Recall dapat dihitung menggunakan rumus sebagai berikut.

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{319}{319+59} = \frac{319}{378} = 0.8439 = 84.39\%$$

Akurasi merupakan persentase prediksi yang benar dari semua pengamatan. Akurasi dapat dihitung menggunakan rumus sebagai berikut.

$$\begin{aligned} \text{Akurasi} &= \frac{TP+TN}{TP+FP+TN+FN} = \frac{319+197}{319+81+197+59} = \frac{516}{656} \\ &= 0.786585365 = 78.6585365\% \end{aligned}$$

Sehingga diperoleh nilai presisi 79.75%, recall 84.39%, dan akurasi 78.6585365%.

Bab IV Penutup

IV.1 Kesimpulan

Penelitian ini bertujuan untuk melakukan analisis dan klasifikasi kualitas air yang dapat di konsumsi berdasarkan parameter yang terdapat dalam database. Exploratory data analysis membantu untuk mempermudah analisis data. Hasil yang diperoleh dari penelitian ini dapat disimpulkan bahwa algoritma *Random Forest Classifier* memiliki akurasi yang paling baik dibandingkan dengan algoritma yang lainnya. Sistem identifikasi kualitas air minum memiliki akurasi sebesar 78% dengan menggunakan algoritma *Random Forest Classifier*.

IV.2 Saran

Beberapa saran yang penyusun ingin sampaikan yaitu:

- Untuk Para Mentor Orbit Future Academy, kami berharap agar selalu sabar dalam menghadapi mahasiswa selama proses belajar mengajar di kelas dan tetap selalu menjalin komunikasi dengan para mentee agar mahasiswa dengan mudah mendapatkan informasi mengenai jalannya pembelajaran.
- Untuk Mitra Orbit Future Academy, kami sangat berterima kasih karena sudah membuka program ini khususnya untuk mahasiswa IT. Kedepannya semoga dalam pelaksanaan pembelajaran lebih interaktif dan menciptakan inovasi pembelajaran yang mudah dipahami oleh mahasiswa bukan IT.

Bab V Referensi

- [1] Arnomo, R. A., Saptomo, W. L. Y., & Harsadi, P. (2018). Implementasi Algoritma K-Nearest Neighbor Untuk Identifikasi Kualitas Air (Studi Kasus : Pdam Kota Surakarta). *Jurnal Teknologi Informasi Dan Komunikasi (TIKomSiN)*, 6(1), 1–5. <https://doi.org/10.30646/tikomsin.v6i1.345>
- [2] Harahap, E. H., Muflikhah, L., & Rahayudi, B. (2018). Implementasi Algoritma Support Vector Machine (SVM) Untuk Penentuan Seleksi Atlet Pencak Silat. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 2(10), 3843–3848.
- [3] Putra, M. I. P. (2019). *Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Kanker Payudara Tugas Akhir diajukan untuk memenuhi salah satu syarat memperoleh gelar sarjana dari Program Studi S1 Ilmu Komputasi M Ikhsan Perdana Putra Program Studi Sarjana Ilmu Komputasi Fakultas Informatika Universitas Telkom Bandung*. 6(1), 2431–2441.
- [4] Riyantoko, P. A. (2021). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. *Seminar Nasional Sains Data 2021 (SENADA 2021)*, 2021(Senada), 12–18.
- [5] Siburian, V., & Mulyana, I. (2021). Prediksi Harga Ponsel menggunakan Metode Random Forest. *Jurnal Komputer Terapan*, 7(1), 24–32.
- [6] 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*. 4(1), 312–319.

Bab VI Lampiran A. TOR

TERM OF REFERENCE (TOR) STUDI INDEPENDEN BERSERTIFIKAT FOUNDATION OF AI AND LIFE SKILLS FOR GEN-Z DI ORBIT FUTURE ACADEMY

A. Rincian Program

Foundation of AI and Life Skills for Gen-Z adalah program pelatihan *Artificial Intelligence* (AI) daring yang bertujuan untuk memperkenalkan teknologi dan perangkat AI kepada pelajar, sehingga diharapkan mereka dapat mengembangkan produk AI yang memiliki dampak sosial. Program ini berfokus pada komponen utama AI, seperti Data Science (DS), Natural Language Processing (NLP), dan Computer Vision (CV). Selain keterampilan AI, pelajar juga akan mendapat *life skills* yang bermanfaat untuk mencari atau menciptakan lapangan kerja.

B. Tujuan Program

Tujuan yang diharapkan setelah peserta mengikuti program ini:

1. Memiliki wawasan tentang AI dan perkembangannya.
2. Mampu merancang dan mengimplementasikan AI Project Cycle.
3. Mampu menggunakan bahasa pemrograman Python untuk mengembangkan aplikasi AI.
4. Mampu mengembangkan salah satu dari 3 domain AI (DS, NLP, dan CV) hingga tahap *deployment*.
5. Mampu menggunakan *soft skills* dan *hard skills* dalam dunia industri dan lingkungan perusahaan.
6. Mampu mengaplikasikan kiat-kiat yang dibutuhkan seorang wirausahawan yang bergerak di bidang *start-up* dalam mentransformasikan ide ke dalam bentuk produk/jasa sehingga dapat menciptakan peluang bisnis yang terus berinovasi, berevolusi, dan berkelanjutan.

C. Jadwal dan Tempat Pelaksanaan

Jadwal pelaksanaan tertera dalam tabel berikut:

Pukul (WIB)	Durasi (jam)	Aktivitas
08.00 s.d. 11.30	3.5	Kelas Sesi Pagi
13.00 s.d. 16.30	3.5	Kelas Sesi Siang
16.30 s.d. 17.30	1	<i>Self-Study</i>

Kelas akan diselenggarakan secara daring melalui aplikasi *video conference*.

D. Peserta

Peserta program ini adalah mahasiswa yang berasal dari Perguruan Tinggi di bawah Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia.

E. Uraian Tugas Peserta

Selama mengikuti program ini, peserta diharuskan:

1. Mengikuti program dari awal hingga selesai.
2. Mematuhi aturan program.
3. Mematuhi aturan kelas yang dibuat bersama *homeroom* atau *domain coach*.
4. Mengikuti kelas dengan presensi minimal 85%.
5. Membuat laporan harian dan mingguan di *website* Kampus Merdeka.
6. Menyelesaikan Proyek Akhir (PA) beserta laporannya.

Homeroom Coach,

Kendal, 21 Februari 2022

Peserta Program,



Angel Metanosa Afinda, S.Kom.
2201043

Sukma Nur Savitri
A410190137

Bab VII Lampiran B. Log Activity

Minggu/Tgl	Kegiatan	Hasil
Minggu 1/25-28 April 2022	Mencari referensi dan menentukan domain PA dan keluaran dari PA.	<ul style="list-style-type: none"> • Domain <i>Data Science</i> • Keluaran <i>Paper</i>
Minggu 2/9-13 Mei 2022	Mencari referensi judul PA	<ul style="list-style-type: none"> • Klasifikasi judul jurnal menggunakan algoritma naïve bayes
9-13 Mei 2022	Berdiskusi dan konsultasi terkait judul paper	<ul style="list-style-type: none"> • Mendapatkan judul paper
16-20 Mei 2022	Berdiskusi mencari dataset	<ul style="list-style-type: none"> • Dataset yang digunakan PA
Minggu 5/30 Mei-3 Juni 2022	Berdiskusi dan konsultasi terkait codingan Berdiskusi terkait pembuatan paper	<ul style="list-style-type: none"> • Merevisi codingan
14 Juni 2022	Bimbingan progress Proyek Akhir	<ul style="list-style-type: none"> • Merevisi Codingan dan <i>Paper</i>
15 Juni 2022	Bimbingan revisi codingan	<ul style="list-style-type: none"> • Mendapat judul baru • Menyempurnakan codingan • Menyusun jurnal
16 Juni 2022	Bimbingan jurnal	<ul style="list-style-type: none"> • Revisi jurnal dan menyusun laporan akhir

Bab VIII Lampiran C. Jurnal

Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma *Machine Learning*

Lidya Savitri¹, Muhammad Rosul Wahidin², Sukma Nur Savitri³, Wafa Sandwi Mustofa⁴

¹Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu

²Program Studi Administrasi Publik, Fakultas Ilmu Sosial dan Politik, Universitas Yudharta Pasuruan

³Program Studi Pendidikan Matematika, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Muhammadiyah Surakarta

⁴Program Studi Pendidikan Matematika, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Muhammadiyah Surakarta

ARTICLE INFO

Article history:

Diterima xx-xx-xx

Diperbaiki xx-xx-xx

Disetujui xx-xx-xx

Kata Kunci:

Klasifikasi, Air, *Machine Learning*

ABSTRAK

Air Minum merupakan air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum. Berdasarkan Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/MENKES/PER/IV/2010 tentang Persyaratan Kualitas Air Minum terdapat pengertian mengenai Air Minum yaitu air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum. Namun, mayoritas masyarakat masih belum mengetahui air yang dapat diminum atau tidak dapat diminum. Artikel ini menjelaskan klasifikasi data sampel air menerapkan algoritma *machine learning*. Tujuan penelitian ini untuk analisis dan klasifikasi kualitas air minum dengan harapan bisa memberikan kontribusi kepada masyarakat sekitar untuk mengetahui kualitas air minum yang baik untuk dikonsumsi.

ABSTRACT

Keywords:

Classification, water, Machine Learning

Drinking water is water that has been processed or without processing that meets health requirements and can be drunk directly. Based on the Regulation of the Minister of Health of the Republic of Indonesia Number 492/MENKES/PER/IV/2010 concerning Drinking Water Quality Requirements, there is an understanding of drinking water, namely through processing or without processing that meets health requirements and can be drunk directly. However, the majority of people still do not know potable or non-drinkable water. This article describes the classification of water sample data using machine learning algorithms. The purpose of this study is to analyze and classify water quality in the hope of contributing to the surrounding community to determine the quality of water that is good for consumption.

1. Pendahuluan

Air bersih merupakan kebutuhan dasar bagi manusia, sehingga ketersediaannya amatlah penting. Air dimanfaatkan dalam kesehariannya tidak hanya terbatas untuk keperluan rumah tangga, tetapi juga untuk fasilitas umum, sosial dan ekonomi. Kebutuhan air bersih terus meningkat seiring dengan perkembangan populasi manusia. Kebutuhan dan permintaan air bersih akan terus meningkat dengan adanya pertumbuhan penduduk, terjadi pergerakan dinamik dalam masyarakat baik dalam segi kepadatan, sosial maupun ekonomi. Hal tersebut membuat manusia terikat dengan fungsi air. Akan tetapi tidak semua zat-zat mineral yang terkandung dalam air dapat dikonsumsi dengan layak.

Air rentan terkena oleh bakteri yang sangat berbahaya bagi sistem pencernaan manusia. Hal ini air terbagi menjadi dua yaitu air yang layak konsumsi dan tidak layak konsumsi. Berdasarkan Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/MENKES/PER/IV/2010 tentang Persyaratan Kualitas Air Minum terdapat pengertian mengenai Air Minum yaitu air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum [1]. Identifikasi dini terhadap produk air dari sumber air baku sangat diperlukan untuk mengetahui faktor-faktor yang mempengaruhi kualitas air yang layak konsumsi bagi masyarakat.

Krisis air bersih sedang melanda berbagai negara di dunia dengan hanya 1% jumlah air bersih yang dapat dikonsumsi oleh manusia. Jumlah yang sangat kecil untuk jumlah air bersih yang baik, menyebabkan air bersih susah diakses oleh penduduk. Sebanyak 663 juta penduduk yang bersumber dari data WHO menunjukkan bahwa mereka susah untuk mengakses air bersih [2]. Berdasarkan data UNESCO, pada tahun 2025 diprediksi bahwa dua pertiga dari jumlah penduduk dunia akan tinggal di daerah yang kekurangan air bersih [2]. Posisi *World Water Assment Programme* (WWAP) dibawah UNESCO sudah melakukan prediksi akan kondisi air bersih untuk beberapa tahun kedepan. Pada kondisi tertentu, sebagai gambaran kebutuhan akan air dalam kehidupan sehari-hari tidak kurang dari 85% air bersih berubah menjadi air limbah. Setiap orang bisa menggunakan hingga 100 liter air perhari untuk memenuhi kebutuhan hidupnya [2].

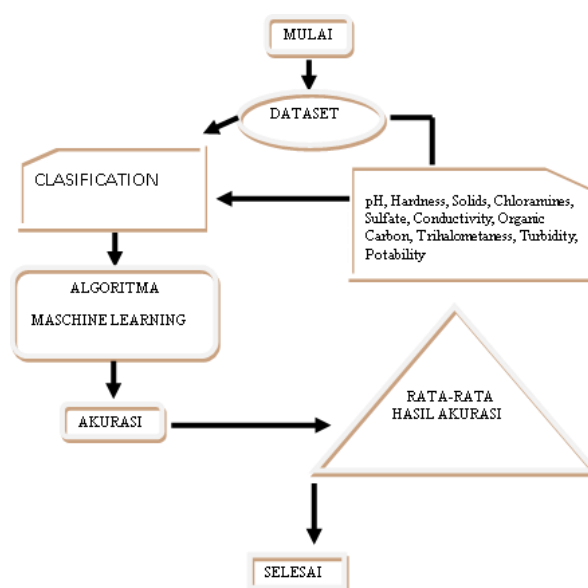
Semakin hari juga tingkat kualitas dari air tersebut semakin menurun. Jika hal itu terus terjadi kondisi seperti ini sangat fatal dan dapat mengakibatkan gangguan pada pencernaan manusia dan seluruh makhluk hidup yang membutuhkan air tersebut. Upaya dilakukan dalam menjaga kualitas air tersebut dengan melakukan pengolahan dengan cara pengawasan di daerah air atau lingkungan sumber daya air. Hal ini berguna agar sumber air tetap terjaga kelestariannya dan dapat menghasilkan kualitas air dengan standart air bersih yang

layak dikonsumsi oleh manusia. Kualitas air bersih yang dimaksud yaitu melakukan pengukuran kondisi terhadap air berdasarkan karakteristik pada fisik air, karakteristik kimiawi dan biologisnya. Pengukuran tersebut untuk mengukur kualitas air tersebut memenuhi syarat yang layak dikonsumsi atau tidak dapat mengetahuinya melalui zat-zat dan mineral yang terkandung di dalam air minum. *Machine learning* lebih memudahkan untuk proses implementasi. *Machine Learning* merupakan salah satu cabang dari ilmu Kecerdasan Buatan, khususnya yang mempelajari tentang bagaimana komputer mampu belajar dari data untuk meningkatkan kecerdasannya.

Penggambaran kualitas air biasanya digambarkan dalam bentuk parameter dan variabel. Beberapa parameter bermacam-macam digunakan sebagai dasar untuk menentukan model seperti apa *paper* ini disusun.

Parameter-parameter tersebut dapat digunakan untuk membuat permodelan klasifikasi dan memprediksi nilai parameter yang lainnya, yang termuat dalam *paper*. Dalam proses memprediksi data dapat digunakan beberapa metode klasifikasi baik secara manual maupun komputasional dengan memanfaatkan *machine learning*. Beberapa algoritma pemodelan yang digunakan pada penelitian ini adalah *Logistic regression*, *SVM*, *Random Forest Classifier*, *KNN*, dan *XGB Classifier*. Penelitian ini, akan menggunakan *semi-supervised machine learning* dengan menggunakan *library python* yang digunakan untuk mengolah data tabular. Metode tersebut digunakan untuk melakukan analisis dan membantu untuk melakukan pre-prosesing data dalam melakukan prediksi dan klasifikasi pada kualitas air minum.

2. Metode Penelitian



Gambar 2.1 Flowchart Proses Analisis Data menggunakan Machine Learning

Bagian ini akan menjelaskan terkait *exploratory data analysis* yang dilakukan.

2.1 Dataset

Penelitian ini merupakan penelitian sederhana dengan menggunakan data yang bersumber dari Kaggle yang bernama *water potability* yang berformat csv yang di selesaikan dengan menggunakan metode *Machine Learning*. Pada data tersebut terdapat sepuluh parameter diantaranya sebagai berikut:

a. pH

pH (*Power of Hydrogen*) adalah skala yang digunakan untuk menyatakan tingkat keasaman atau kebasaan yang dimiliki oleh suatu larutan. Skala dari pH terdiri dari angka 1 hingga 14.

b. Hardness

Hardness adalah kandungan mineral-mineral tertentu di dalam air. Umumnya ion kalsium (Ca) dan magnesium (Mg) dalam bentuk garam karbonat.

c. Solids

Istilah untuk menandakan jumlah padatan terlarut atau konsentrasi jumlah ion kation (bermuatan positif) dan anion (bermuatan negatif) di dalam air.

d. Chloramines

Sebuah kompleks kimia yang terdiri dari klorin dan amonia.

e. Sulfate

Kandungan yang merupakan foaming agent (mampu menimbulkan busa) dan biasa dipakai pada produk seperti pembersih wajah, sampo, dan pasta gigi.

f. Conductivity

Konduktivitas adalah ukuran kemudahan di mana muatan listrik atau panas dapat melewati suatu bahan.

g. Organic Carbon

Karbon memiliki TOC yang digunakan untuk mengetahui jumlah total carbon dalam air murni.

h. Trihalomethanes

Trihalomethanes (THMs) adalah hasil reaksi antara klorin yang digunakan untuk mendisinfeksi air keran dan bahan organik alami di dalam air. Pada tingkat tinggi, THMs telah dikaitkan dengan efek kesehatan negatif seperti kanker dan hasil reproduksi yang merugikan.

i. Turbidity

Turbidity (kekuruhan) adalah ukuran kejernihan relatif suatu cairan.

j. Potability

Indikator air yang layak konsumsi dan tidak konsumsi.

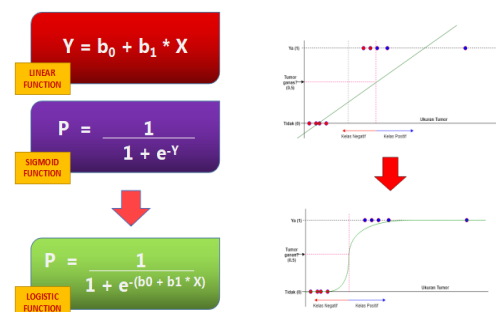
2.2 Clasification

Classification dalam *data science* berarti proses memprediksi kelas atau kategori data dengan memanfaatkan nilai yang ada pada data. Algoritma *machine learning* sendiri dibagi menjadi dua, yaitu *supervised* dan *unsupervised learning*. *Classification* termasuk dalam algoritma *supervised learning*, selain *classification* terdapat *regression* dan *forecasting*. Algoritma yang digunakan dalam *classification* sendiri sangat beragam. Anda bisa memilih antara *logistic regression*, *random forest*, dan lain-lain. Proses *classification* pada dasarnya dilakukan agar analisis data menjadi lebih mudah dan tentunya memberikan hasil yang akurat. Agar bisa memberikan suatu informasi yang bermanfaat, data memang memerlukan proses panjang.

2.3 Algoritma Machine Learning

2.3.1 Logistic Regression

Logistic regression adalah jenis analisis statistik yang sering digunakan data analyst untuk pemodelan prediktif. Dalam pendekatan analitik ini, variabel dependennya terbatas atau kategoris, bisa berupa A atau B (regresi biner) atau berbagai opsi hingga A, B, C atau D (regresi multinomial). Jenis analisis statistik digunakan dalam *software* statistik untuk memahami hubungan antara variabel dependen dan satu atau lebih variabel independen dengan memperkirakan probabilitas. Jenis analisis ini dapat membantu Anda memprediksi kemungkinan.



Gambar 2.1 Rumus dan Grafik *Logistic Regression*

2.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma pembelajaran *machine* terawasi yang dapat digunakan untuk tantangan klasifikasi atau regresi. Namun, sebagian besar digunakan dalam masalah klasifikasi. Dalam algoritma SVM, kami memplot setiap item data sebagai titik dalam ruang n-dimensi (n adalah sejumlah fitur yang Anda miliki) dengan nilai setiap fitur menjadi nilai koordinat tertentu.

2.3.3 Random Forest Classifier

Random Forest adalah pengklasifikasi yang berisi sejumlah pohon keputusan pada berbagai subset dari dataset yang diberikan dan mengambil rata-rata untuk meningkatkan akurasi prediksi dari dataset itu.

2.3.4 K-Nearest Neighbor (KNN)

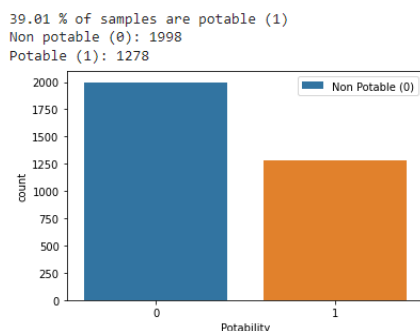
K-Nearest Neighbor adalah salah satu algoritma *Machine Learning* yang paling sederhana berdasarkan teknik *Supervised Learning*. Algoritma K-NN mengasumsikan kesamaan antara kasus/data baru dengan kasus yang tersedia dan memasukkan kasus baru ke dalam kategori yang paling mirip dengan kategori yang tersedia. Algoritma K-NN menyimpan semua data yang tersedia dan mengklasifikasikan titik data baru berdasarkan kesamaan. Artinya ketika data baru muncul maka dapat dengan mudah diklasifikasikan ke dalam kategori *well suite* dengan menggunakan algoritma K-NN.

2.3.5 XGBoost Classifier

XGBoost adalah implementasi dari pohon keputusan yang didorong gradien yang dirancang untuk kecepatan dan kinerja. *XGBoost* adalah algoritma peningkatan gradien ekstrim. Dan itu berarti ini adalah algoritma pembelajaran mesin yang besar dengan banyak bagian. *XGBoost* bekerja dengan kumpulan data yang besar dan rumit. *XGBoost* adalah teknik pemodelan *ensemble*.

3. Hasil dan Pembahasan

Hasil dari *exploratory data analysis* total data sebanyak 3276 data dan 10 kolom parameter.



Gambar 3.1 Diagram kualitas air minum berdasarkan database

Berdasarkan hasil diagram di atas dari 3276 data bahwa terdapat 1998 *non potable* atau air tidak dikonsumsi dan 1278 *potable* atau air dapat dikonsumsi.

Proses klasifikasi kualitas air minum melewati proses modelling yang dilakukan dengan menggunakan lima algoritma *machine learning* menghasilkan nilai yang berbeda antara algoritma satu dengan yang lain. **Tabel 3.1** akan menunjukkan hasil dari proses modelling lima algoritma tersebut

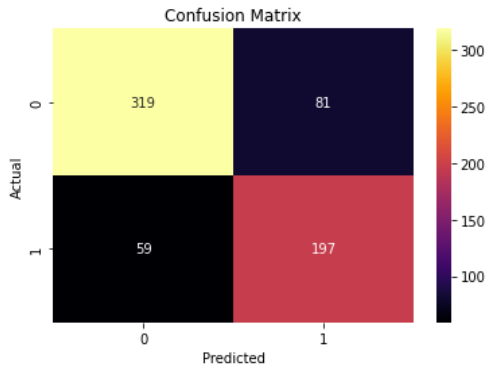
Tabel 3.1 Hasil Klasifikasi

Classifier	Class	Precision	Recall	f1-score	Akurasi
Logistic Regression	0	0.64	0.51	0.56	0.52
	1	0.42	0.55	0.47	
SVM	0	0.75	0.65	0.70	0.66
	1	0.55	0.66	0.60	
XGB Classifier	0	0.87	0.69	0.77	0.75
	1	0.64	0.84	0.73	
Random Forest	0	0.85	0.79	0.81	0.78
	1	0.70	0.78	0.74	
KNN	0	0.71	0.63	0.67	0.62
	1	0.51	0.61	0.56	

Berdasarkan hasil diatas bahwa pada *classifier logistic regression* memiliki *precision* kelas 0 adalah 64% dan kelas 1 adalah 42%, *recall* kelas 0 adalah 51% dan kelas 1 adalah 55%, *f1-score* kelas 0 adalah 56% dan kelas 1 adalah 47%, serta akurasinya sebesar 52%. Pada *classifier SVM* memiliki *precision* kelas 0 adalah 75% dan kelas 1 adalah 55%, *recall* kelas 0 adalah 65% dan kelas 1 66%, *f1-score* kelas 0 adalah 70% dan kelas 1 adalah 60%, serta akurasinya sebesar 66%. Pada *XGB classifier* memiliki *precision* kelas 0 adalah 87% dan kelas 1 adalah 64%, *recall* kelas 0 adalah 69% dan kelas 1 84%, *f1-score* kelas 0 adalah 77% dan kelas 1 adalah 73%, serta akurasinya sebesar 74%. Pada *random forest classifier* memiliki *precision* kelas 0 adalah 85% dan kelas 1 adalah 70%, *recall* kelas 0 adalah 79% dan kelas 1 78%, *f1-score* kelas 0 adalah 81% dan kelas 1 adalah 74%, serta akurasinya sebesar 78%. Pada *classifier KNN* memiliki *precision* kelas 0 adalah 71% dan kelas 1 adalah 51%, *recall* kelas 0 adalah 63% dan kelas 1 61%, *f1-score* kelas 0 adalah 67% dan kelas 1 adalah 56%, serta akurasinya sebesar 62%.

Tingkat akurasi yang dihasilkan setiap model berbeda. Berdasarkan urutan model dengan tingkat akurasi terendah ke tertinggi dapat dilihat sebagai berikut *logistic regression*, *K-*

Nearest Neighbor, Support Vector Machine, XGB Classifier, dan Random Forest Classifier. Oleh karena itu, dapat disimpulkan bahwa *Random Forest Classifier* memiliki akurasi tertinggi yaitu 78%.



Gambar 3.2 Confusion Matrix Random Forest Classifier

Berdasarkan Gambar 3.2 akan diketahui nilai presisi, recall, dan akurasi pada *random forest classifier*. Nilai yang dihasilkan gambar diatas dapat diartika bahwa nilai 319 yang berarti *True Positive* (TP), nilai 197 yang berarti *True Negative* (TN), nilai 81 berarti *False Positif* (FP), dan nilai 59 berarti *False Negative* (FN).

Presisi merupakan persentase kasus yang diprediksi positif yang ternyata benar. Presisi dapat dihitung dengan menggunakan rumus sebagai berikut.

$$\text{Presisi} = \frac{TP}{TP+FP} = \frac{319}{319+81} = \frac{319}{400} = 0.7975 = 79.75\%$$

Recall digunakan untuk mengukur pecahan kasus positif yang diidentifikasi dengan benar. Recall dapat dihitung menggunakan rumus sebagai berikut.

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{319}{319+59} = \frac{319}{378} = 0.8439 = 84.39\%$$

Akurasi merupakan persentase prediksi yang benar dari semua pengamatan. Akurasi dapat dihitung menggunakan rumus sebagai berikut.

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{319+197}{319+81+197+59} = \frac{516}{656} = 0.786585365 = 78.6585365\%$$

Sehingga diperoleh nilai presisi 79.75%, recall 84.39%, dan akurasi 78.6585365%.

4. Kesimpulan

Penelitian ini bertujuan untuk melakukan analisis dan klasifikasi kualitas air yang dapat di konsumsi berdasarkan parameter yang terdapat dalam *database*. *Exploratory data analysis* membantu untuk mempermudah analisis data. Hasil yang diperoleh dari penelitian ini dapat disimpulkan bahwa algoritma *Random Forest Classifier* memiliki akurasi yang paling baik dibandingkan dengan algoritma yang lainnya. Sistem identifikasi kualitas air minum memiliki akurasi sebesar

78% dengan menggunakan algoritma *Random Forest Classifier*.

Namun terlepas dari itu semua, kami menyadari bahwa penelitian ini masih jauh dari kata sempurna, beberapa kendala yang kami alami semisal iklim komunikasi yang masih belum optimal, kesadaran atas apa yang seharusnya secara proaktif harus dilakukan, dan tentunya dengan pengetahuan yang masih dibidang terbatas membuat kami kesusahan dalam menyempurnakan penelitian ini.

Ucapan Terima Kasih

Tidak dapat disangkal bahwa proses mengerjakan penelitian ini bisa berjalan dengan lancar oleh karena berkat upaya partisipatif baik dari teman sekelompok dan coach pendamping project akhir. Maka kami mengucapkan terima kasih sebesar-besarnya jajaran mitra Orbit Future Academy, khususnya kepada coach Angel atas kesediannya memberikan berbagai bantuan dan pengawalan kepada kami yang dinilai cukup intens. Pada akhirnya puji syukur Alhamdulillahirobbilalamin sehingga kami dapat menyelesaikan *project* akhir ini, sekian terima kasih.

Referensi

- [1] K. Liu, Z. Li, C. Yao, J. Chen, K. Zhang, and M. Saifullah, "Coupling the k-nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting," *Int. J. Sediment Res.*, vol. 31, no. 2, pp. 149–158, 2016.
- [2] Riyantoko, P. A. (2021). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. *Seminar Nasional Sains Data 2021 (SENADA 2021)*, 2021(Senada), 12–18.
- [3] Rochmi, MN. "Akses air bersih masih jauh banget dari target". Diakses dari: <https://beritagar.id/artikel/editorial/hapuskan-perda-penyebab-ekonomi-biaya-tinggi>. 2016.
- [4] UNESCO. "Global Climate Change". Diakses dari: www.unesco.org. 19 Agustus 2021.
- [5] Elysia, V. "Air dan Sanitasi dimana posisi Indonesia". *Seminar Nasional Peran Matematika, Sains, dan Teknologi dalam mencapai tujuan pembangunan berkelanjutan/SDGs, FMIPA Universitas Terbuka*. 2018. Halaman 157-159.
- [6] Artikel Saintif.com. Saintif. Published November 24, 2020. Accessed June 17, 2022. <https://saintif.com/ph-adalah/>

[7]	Water Hardness Tester YD300 - Ukur Kesadahan Air. Digilifeweb.com. Published 2022. Accessed June 17, 2022. http://digilifeweb.com/Water-Hardness-Tester-YD300		https://www.analyticsvidhya.com/blog/2017/09/understanding-support-vector-machine-example-code/ (accessed 2022 -06 -17).
[8]	chloramines – Termwiki, millions of terms defined by people like you https://id.termwiki.com/ID/chloramines (accessed 2022 -06 -17).	[19]	Machine Learning Random Forest Algorithm - Javatpoint https://www.javatpoint.com/machine-learning-random-forest-algorithm (accessed 2022 -06 -17).
[9]	https://journal.sociolla.com/bjglossary/sulfate (accessed 2022 -06 -17).	[20]	K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed 2022 -06 -17).
[10]	Conductivity https://www.lehigh.edu/~amb4/wbi/kwardlow/conductivity.htm#:~:text=Conductivity%20is%20the%20measure%20of,metals%2C%20semiconductors%2C%20and%20insulators (accessed 2022 -06 -17).	[21]	Agrawal, K. What is the XGboost classifier algorithm? XGBoost classifier is a Machine learning algorithm that is applied. https://www.linkedin.com/pulse/xgboost-classifier-algorithm-machine-learning-kavya-kumar#:~:text=What%20is%20the%20XGboost%20classifier,an%20extreme%20gradient%20boost%20algorithm (accessed 2022 -06 -17).
[11]	Organic Carbon		
[12]	https://serc.carleton.edu/microbelife/research_methods/biogeochemical/organic_carbon.html (accessed 2022 -06 -17).		
[13]	Hood, E. Tap Water and Trihalomethanes: Flow of Concerns Continues. <i>Environmental Health Perspectives</i> 2005, 113 (7), A474.		
[14]	Turbidity and Water U.S. Geological Survey https://www.usgs.gov/special-topics/water-science-school/science/turbidity-and-water#:~:text=Turbidity%20is%20the%20measure%20of,light%2C%20the%20higher%20the%20turbidity (accessed 2022 -06 -17).		
[15]	Apa Itu Classification dalam Data Science? https://algorit.ma/blog/classification-adalah-2022/ (accessed 2022 -06 -17).		
[16]	3 Tipe Logistic Regression yang Wajib Diketahui Data Analyst https://algorit.ma/blog/logistic-regression-adalah-2022/ (accessed 2022 -06 -17).		
[17]	SVM Support Vector Machine Algorithm in Machine Learning		
[18]			