Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma *Machine Learning*

Lidya Savitri[1], Muhammad Rosul Wahidin[2], Sukma Nur Savitri[3], Wafa Sandwi Mustofa[4]

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

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

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

[4]Program 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.

Keywords:
Classification, water, Machine Learning

ABSTRACT

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.

Email: lidyasavitri013@gmail.com, rosulwahidi@gmail.com, sukmavivi71@gmail.com, wafamusthofa@gmail.com

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 dikomsumsi 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.

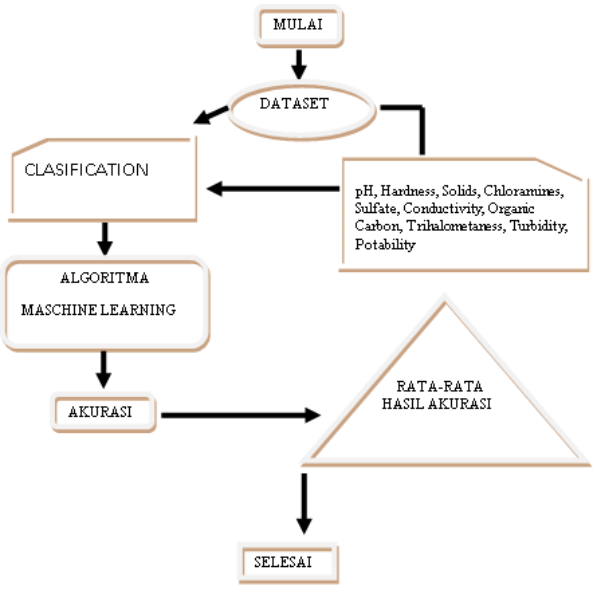
Krisis air bersih sedang melanda berbagai negara di dunia dengan hanya 1% jumlah air bersih yang dapat dikonsumsi oleh manusia. Jumlah yang sangat kecil untuk jumlah air bersih yang baik, menyebabkan air bersih susah diakses oleh penduduk. Sebanyak 663 juta penduduk yang bersumber dari data WHO menunjukkan 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 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 [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 dikomsumsi 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 dikomsumsi atau tidak dapat mengetahuinya melalui zat-zat dan mineral yang terkandung di dalam air minum. *Machine learning* lebih memumudah 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.

Pengambaran 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 manfaatkan *machine learning.* Beberapa algoritma pemodelan yang digunakan pada penelitian ini adalah *Logistic regression, SVM, Random Forest Classifer, 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

Email: lidyasavitri013@gmail.com, rosulwahidi@gmail.com, sukmavivi71@gmail.com, wafamusthofa@gmail.com

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 tersesbut 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.

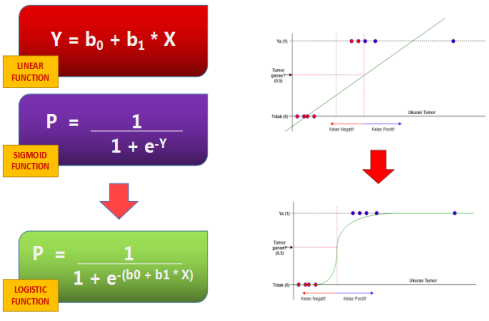
2.2 *Clasification*

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*

Email: lidyasavitri013@gmail.com, rosulwahidi@gmail.com, sukmavivi71@gmail.com, wafamusthofa@gmail.com

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)

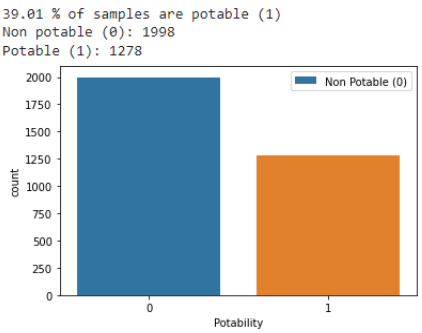
K-Nearest Neighbor adalah salah satu algoritma *Machine Learning* yang paling sederhana berdasarkan teknik *Supervised Learning*. Algoritma K-NN mengasumsikan kesamaan antara kasus/data baru dengan kasus yang tersedia dan memasukkan kasus baru ke dalam kategori yang paling mirip dengan kategori yang 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 Clasifier*

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

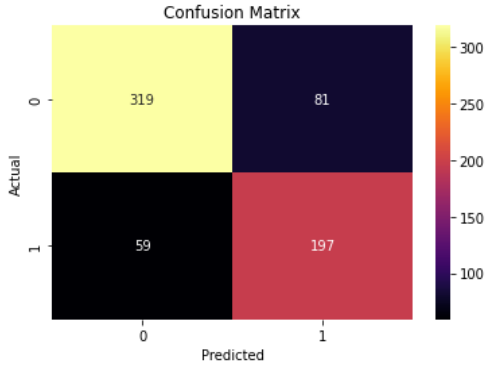
Classifier	Class	Precision	Recall	f1-score	Akurasi
Logistic Regression	0	0.64	0.51	0.56	0.52
	1	0.42	0.55	0.47	
SVM	0	0.75	0.65	0.70	0.66
	1	0.55	0.66	0.60	
XGB Classifier	0	0.87	0.69	0.77	0.75
	1	0.64	**0.84**	0.73	
Random Forest	0	**0.85**	0.79	**0.81**	**0.78**
	1	0.70	0.78	0.74	
KNN	0	0.71	0.63	0.67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

Email: lidyasavitri013@gmail.com, rosulwahidi@gmail.com, sukmavivi71@gmail.com, wafamusthofa@gmail.com

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 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.

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

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 dibilang 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 Alhamdulillahhirobbilalamin sehiingga 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/

Email: lidyasavitri013@gmail.com, rosulwahidi@gmail.com, sukmavivi71@gmail.com, wafamusthofa@gmail.com

[7]	Water Hardness Tester YD300 - Ukur Kesadahan Air. Digilifeweb.com. Published 2022. Accessed June 17, 2022. http://digilifeweb.com/Water-Hardness-Tester-YD300
[8]	chloramines – Termwiki, millions of terms defined by people like you https://id.termwiki.com/ID/chloramines (accessed 2022 -06 -17).
[9]	https://journal.sociolla.com/bjglossary/sulfate (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).
[12]	Organic Carbon 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. *Environmental Health Perspectives* 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).
[16]	Apa Itu Classification dalam Data Science? https://algorit.ma/blog/classification-adalah-2022/ (accessed 2022 -06 -17).
[17]	3 Tipe Logistic Regression yang Wajib Diketahui Data Analyst https://algorit.ma/blog/logistic-regression-adalah-2022/ (accessed 2022 -06 -17).
[18]	SVM | Support Vector Machine Algorithm in Machine Learning https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (accessed 2022 -06 -17).
[19]	Machine Learning Random Forest Algorithm - Javatpoint https://www.javatpoint.com/machine-learning-random-forest-algorithm (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).
[21]	Agrawal, K. What is the XGboost classifier algorithm? XGBoost classifier is a Machine learning algorithm that is applie. 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).

Email: lidyasavitri013@gmail.com, rosulwahidi@gmail.com, sukmavivi71@gmail.com, wafamusthofa@gmail.com