Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning

Prismahardi Aji Riyantoko[1], Tresna Maulana Fahrudin[2], Kartika Maulida Hindrayani[3]
[1,2,3]*Sains Data, UPN "Veteran" Jawa Timur*
[2]tresna.maulana.ds@upnjatim.ac.id
[3]kartika.ds@upnjatim.ac.id
Corresponding author email: prismahardi.aji.ds@upnjatim.ac.id

Abstract: *Water is one of the minerals that are very important for the needs of life, especially human survival. Water quality can be detected based on the related mineral parameters contained in it. An indicator that greatly affects water fit for consumption is water potability with a value of 0 which cannot be consumed and a value of 1 can be consumed. This can be forecast using machine learning methods. One of the semi-supervised machine learning methods is the Lucifer Machine Learning Technique. The method develops a library in python that includes exploratory data analysis, skewness correction, and model classification to easily get a high level of accuracy. By using an open database, the data that we collect using the Lucifer Machine Learning Technique gets accurate results. The classification model provides a various levels of accuracy above 50%. The highest accuracy results were obtained from the Random Forest Classifier classification model with an accuracy rate of 72.81%. A simple analysis of the classification model using Lucifer-ML is very easy to understand, therefore this research produces a very simple analysis but makes it easier for readers to understand fundamentally.*
Keywords: *Lucifer, Semi-Supervised Machine Learning, Classification, Simple Analytics*

Abstrak: Air merupakan salah satu mineral yang sangat penting untuk kebutuhan hidup, terutama kelangsungan hidup manusia. Kualitas air dapat dideteksi berdasarkan keterkaitan parameter mineral yang terkandung dalamnya. Indikator yang sangat mempengaruhi suatu air layak dikonsumsi adalah *water potability* dengan nilai 0 tidak dapat dikonsumsi dan nilai 1 dapat dikonsumsi. Hal ini dapat diprediksi dengan menggunakan metode machine learning. Salah satu metode *semi-supervised* machine learning yaitu Teknik Lucifer Machine Learning. Metode tersebut mengembangkan sebuah library pada python yang mencakup *exploratory data analysis, skewness correction,* dan klasifikasi model untuk mendapatkan tingkat akurasi yang tinggi secara mudah. Dengan menggunakan database terbuka, data yang kami oleh dengan menggunakan Teknik Lucifer Machine Learning mendapatkan hasil yang akurat. Model klasifikasi menghasilkan beragam tingkat akurasi diatas 50%. Hasil akurasi tertinggi didapatkan dari model klasifikasi *Random Forest Classifier* dengan tingkat akurasi sebesar 72,81%. Analisis sederhana pada model klasifikasi dengan menggunakan LuciferML sangat mudah dipahami, oleh karena itu dalam penelitian ini menghasilkan analisis yang sangat sederhana tetapi memudahkan pembaca untuk memahami secara fundamental.
Kata kunci: Lucifer, Semi-Supervised Machine Learning, Klasifikasi, Analisis Sederhana

I. PENDAHULUAN

Krisis air bersih sedang melanda berbagai negara di dunia dengan jumlah air bersih hanya 1% yang dapat dikonsumsi oleh manusia. Dengan jumlah yang sangat kecil 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 [1]. Berdasarkan data UNESCO, pada tahun 2025 diprediksi bahwa dua pertiga dari jumlah penduduk dunia akan tinggal di daerah yang kekurangan air bersih [2]. *World Water Asssment Programme* (WWAP) sudah melakukan prediksi akan kondisi air bersih untuk beberapa tahun kedepan, WWAP berada dibawah UNESCO. 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 [3]. Air dan sanitasi sangat berkaitan untuk pengelolaan air bersih, karena fasilitas sanitasi yang layak dapat memenuhi kebutuhan hidup bersih dan sehat yang merupakan elemen penting dalam meningkatkan derajat kesehatan penduduk dunia. Dalam penerapannya air minum yang disarankan untuk bisa langsung diminum dapat melihat kandungannya berdasarkan zat mikrobiologi, kimia fisika dan radio aktif [4-6]. Air dalam tubuh manusia berkisar diantara 50-70% dari seluruh berat badan manusia. Dalam tubuh manusia mengandung air yang terdiri dari 80% kandungan air dalam darah, sebaliknya apabila kekurangan 15% dari berat badan dapat mengakibatkan kematian [7-9].

Pengambaran kualitas air biasanya digambarkan dalam bentuk parameter dan variabel. Beberapa parameter bermacam-macam digunakan sebagai dasar untuk menentukan model seperti pada paper

[10-12]. Berdasarkan parameter tersebut, dalam beberapa penelitian digunakan untuk menentukan prediksi nilai parameter selanjutnya, yang tertuang pada paper [13]. Dalam memprediksi data, digunakan beberapa metode klasifikasi data yang ditentukan secara manual dan komputasional dengn manfaatkan *machine learning.* Beberapa contoh klasifikasi dalam *machine learning* yang bisa digunakan antara lain *Support Vector Machine, Decision Tree, Naïve Bayes,* dan *Artificial Neural Network.* Masing-masing metode klasifikasi data tersebut memiliki kelebihan maupun kekurangan. Metode Regresi Logistik dan *Support Vector Machine* merupakan metode yang digunakan untuk menyelesaikan permasalahan dengan jumlah data yang besar [14-15]. Metode tersebut merupakan bagian dari *Supervised Machine Learning.* Dalam 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.

II. METODE PENELITIAN

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 *semi-supervised machine learning* yaitu metode *Lucifer Machine Learning* [16]. Pada data tersesbut terdapat sepuluh parameter diantaranya sebagai berikut

Tabel 1. Parameter Data Kualitas Air Minum

Parameter	Deskripsi
pH Value	Parameter yang digunakan untuk mengevaluasi keseimbangan asam-basa didalam air
Hardness	Paramter yang digunakan untuk mendeteksi kalsium dan garam magnesium
Total Dissolved Solids	Kemampuan air untuk melarutkan berbagai mineral
Chloramines	Kandungan klorin dan kloramin sebagai disinfektan dalam air
Sulfate	Sulfate kandungan yang ditemukan dalam mineral, tanah, dan batuan
Conductivity	Untuk mengetahui bahwa air murni bukanlah penghantar arus listrik yang baik, melainkan isolator yang baik.
Organic Carbon	Karbon memiliki TOC yang digunakan untuk mengetahui jumlah total carbon dalam air murni
Trihalomethanes	Senyawa kimia yang dapat ditemukan dalam air yang diolah dengan klorin
Turbidity	Kekeruhan air tergantung pada julah zat padat yang ada dalam kondisi tersuspensi
Potability	Indikator air yang layak konsumsi dan tidak konsumsi

Pada penelitian ini terdapat beberapa Langkah untuk mengolah data yang divisualisaikan dengan diagram sebagai berikut

E-ISSN 2808-5841
P-ISSN 2808-7283

Gambar 1. Proses Analisis Data Menggunakan Lucifer Machine Learning

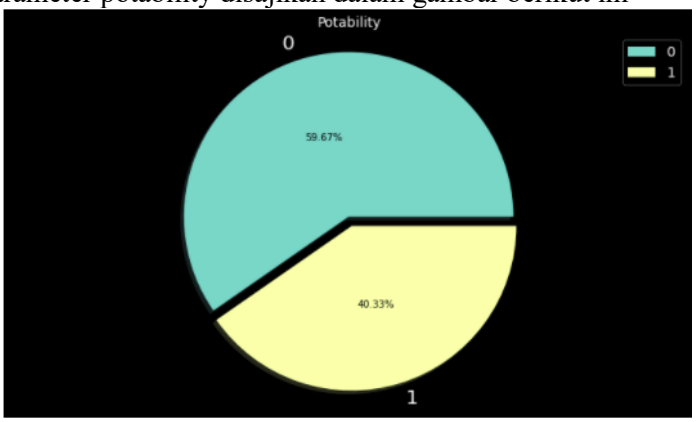
Untuk menyelesaikan metode Lucifer Machine Learning diperlukan langkah-langkah untuk mengolah data, dari proses preprosesing sampai setelah proses pengolahan data. Lucifer Machine Learning merupakan sebuah metode yang terdapat didalam library python. Library pada Python merupakan sebuah kode tambahan yang digunakan untuk menyelesaikan permasalahan tertentu. Hal sangat dimungkinkan untuk setiap orang bisa membuat serta mengembangkan banyak library pada Bahasa Pemrograman Python yang didukung banyak library tidak berbayar alias gratis.

Sebelum memulai pengolahan data, terlebih dahulu dilakukan proses analisis data dengan melakukan proses membaca data atau biasanya yang kita kenal dengan proses *import database.* Proses tersebut dapat dikatakan sebagai preprosesing data. Selanjutnya, proses *exploratory data analysis* yang digunakan untuk memunculkan ukuran dan jumlah data, mean dan standar deviasi, nilai kuartil 1, 2, dan 3, serta nilai maksimum dan minimum. Dalam proses ini juga terdapat visualisasi data berupa *box-plot* untuk menampilkan nilai numerik pada data, *pie-chart, correlation plot, distribution plots,* dan *pair plots.*

Berdasarkan beberapa bentuk visualisasi terutama pada hasil *distribution plots*, pada library LuciferML juga di tambahkan dengan *Skewness Correction.* Proses *skewness* atau ketidaksimetrian dalam distribusi nilai dapat berupa nilai positif, negative, dan nol. Untuk membuat penelitian ini menjadi lebih lengkap, metode klasifikasi yang ada di dalam library tersebut ada banyak, tetapi dalam penelitian ini digunakan metode klasifikasi sebagai berikut *Logistic Regression, Support Vector Machine, K-NN, Decision Trees, Naïve Bayes, Random Forest Classifier, XGBosst Classifier,* dan *Artificial Neural Network.* Pada model klasifikasi tersebut akan dimunculkan nilai akurasi dan standar deviaasi yang didapatkan.

III. HASIL DAN PEMBAHASAN

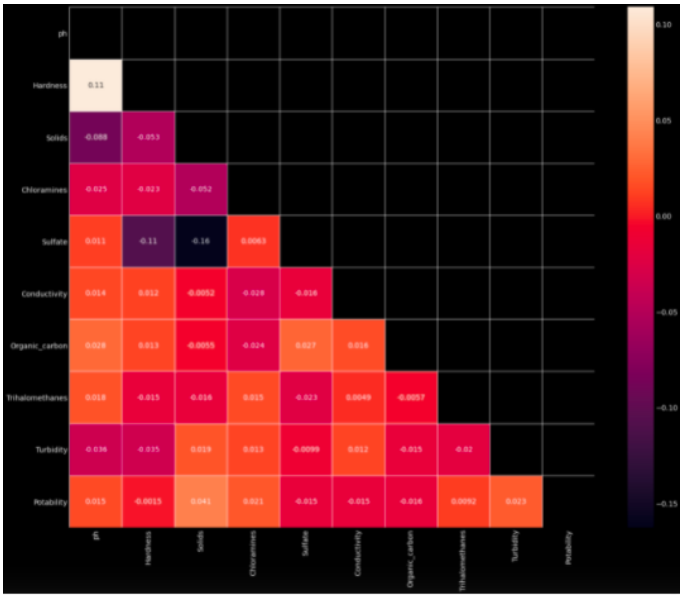
Pada penelitian ini, hasil exploratory data analysis terdapat 3276 data, dan 10 kolom parameter. Hasil penelitian pada parameter potability disajikan dalam gambar berikut ini



Gambar 2. Persentase Kualitas Air Minum berdasarkan Database

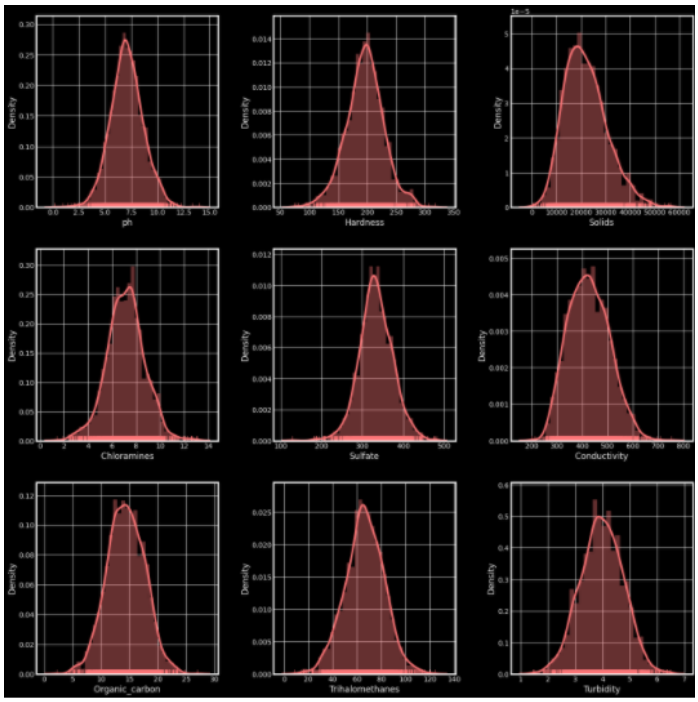
Berdasarkan hasil observasi mengenai persentase kualitas air minum berdasarka database *Water Potability,* dari 3276 data bahwa terdapat 59,67% air tidak dikonsumsi dan 40,33% air dapat dikonsumsi.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 3. Plotting Korelasi antar Parameter

Dari Gambar 3 terdapat tiga parameter yang memiliki hubungan antar parameter yang rendah diantaranya Sulfate-Solids sebesar 0,16 poin dan Sulfate-Hardness 0,11 poin. Selain itu juga terdapat beberapa parameter tidak memiliki korelasi yang erat antar parameter. Dari indicator warna dapat kita baca, bahwa semakin cerah warna yang tertera korelasi antara dua parameter tersebut semakin besar, begitu sebaliknya apabila indicator warna semakin pekat, korelasi antara dua parameter tersebut semakin kecil.



Gambar 4. *Skewness Correction*

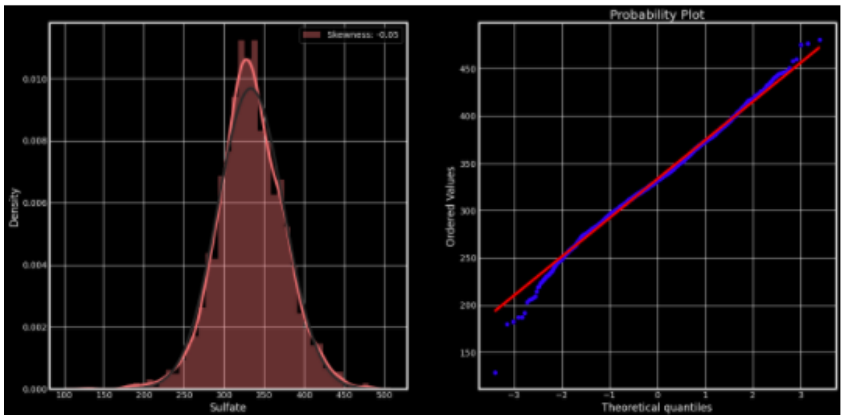
Berdasarkan hasil dari diagram *skewness* tersebut terdapat sembilan parameter yang memiliki nilai positif dan negatif.

Tabel 2. Perbandingan Nilai Skewness, Mean, dan Standar Deviasi sebelum dan sesudah *Skewness Correction*

Parameter	Sebelum			Setelah		
	Skewness	Mean	Standar Deviasi	Skewness	Mean	Standar Deviasi
pH Value	0,0489	7,08	1,57	-1,1353	2,06	0,21
Hardness	-0,0851	195,96	32,62	0,8204	5,26	0,17
Total Dissolved Solids	0,5954	21917,44	8640	-1,2308	9,90	0,44

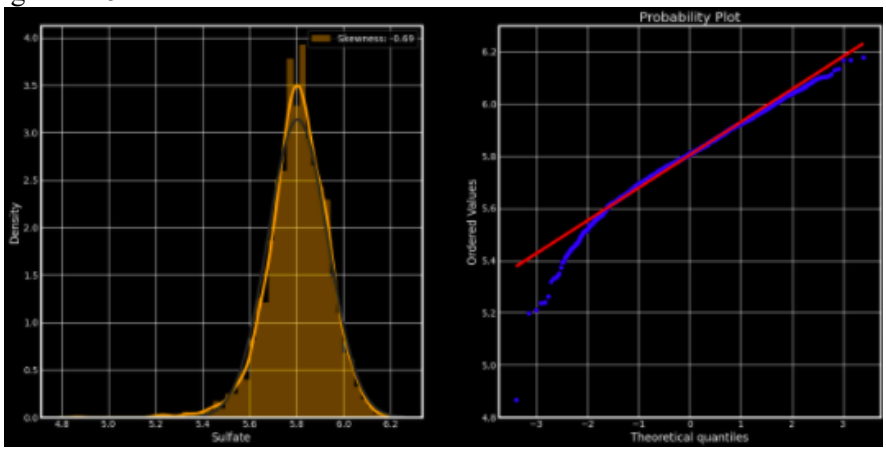
Chloramines	0,0129	7,13	1,58	-0,6533	2,07	0,21
Sulfate	-0,0465	333,22	41,19	-0,6904	5,80	0,12
Conductivity	0,2666	426,56	80,69	-0,1992	6,03	0,19
Organic Carbon	-0,0200	14,35	3,32	-0,8735	2,70	0,23
Trihalomethanes	-0,0513	66,40	16,07	-1,1717	4,17	0,26
Turbidity	-0,0330	3,96	0,78	-0,5344	1,59	0,16

Pada Tabel 2 terdapat data informasi mengenai perbaikan nilai *Skewness.* Pada proses Teknik *skewness correction* berguna untuk menggambarkan bentuk sebuah distribusi atau kurva dan mengetahui pemusatan data yang dapat mewakili suatu data. Sebagai contoh dari hasil Teknik tersebut, diberikan bentuk kurva sebelum dan sesudah Teknik *Skewness Correction* dibuat, misalkan ambil parameter Sulfate dengan kurva sebagai berikut



Gambar 5. Distribusi Plotting Sulfate (Sebelum)

Berdasarkan nilai *Skewness* pada Tabel 2, sebelum dilakukannya perbaikan nilai *Sulfate* -0,0465, mengidikasikan kurva menceng ke kanan, perbedaan tersebut dapat dilihat pada kurva bergaris hitam dan merah pada gambar 5.



Gambar 6. Distribusi Plotting Sulfate (Sesudah)

Setelah diberlakukannya Teknik *Skewness Correction* dapat dilihat pada Gambar 6 dengan melihat tingkat kesimetrian kurva pada indicator kurva bergaris hitam dan kuning. Dapat disimpulkan bahwa dua kurva tersebut saling simetris.

Tabel 3. Model Klasifikasi

Jenis Model	Akurasi	Standard Deviasi
Logistic Regression	51,86%	0,85%
Support Vector Machine	54,37%	2,56%
K-Nearest Neighbor	72,03%	3,30%
Decision Trees	63,28%	4,12%
Naïve Bayes	57,12%	2,85%
Random Forest Classifier	72,81%	2,71%
XGBoost Classifier	69,27%	2,42%
Artificial Neural Network	65,62%	3,04%

Dari Tabel 3 dapat diketahui bahwa tingkat akurasi setiap model berbeda, tetapi tingkat akurasi model diatas 50%. Berdasarkan urutan model terendah ke tertinggi dapat dilihat sebagai berikut *Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Trees, Artificial Neural Network, XGBoost Classifier, K-Nearest Neighbor,* dan *Random Forest Classifier.* Dapat disimpulkan bahwa *Logistic Regression* memiliki tingkat akurasi terendah dengan 51,86% dan *Random Forest Classifier* memiliki nilai akurasi tertinggi dengan 72,81%. Tetapi dapat kita lihat bahwa *K-Nearest Neighbor* juga memiliki tingkat akurasi hampir mendekati model klasifikasi tertinggi terpaut 0,78%.

IV. KESIMPULAN

Pada penelitian ini memiliki tujuan untuk melakukan analisis sederhana pada data kualitas air yang dapat di konsumsi berdasarkan parameter-paramater yang terdapat pada database. Tetapi peneliti berfokus pada data *water potability* dengan parameter acuannya yaitu *Potability* atau kelayakan suatu air untuk dapat dikategorikan konsumsi atau tidak dapat dikonsumsi. Database tersebut kami olah menggunakan metode Lucifer Machine Learning yang merupakan sebuah library program python yang dikembangkan oleh *dark-lucifer* untuk memudahkan melakukan tahapan analisis data yang berdasarkan *exploratory data analysis,* Teknik *skewness correction,* dan beberapa model klasifikasi. 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 tertingi yaitu *Random Forest Classifier* sebesar 72,81%. Hal ini masih jauh dari kategori sempurna, tetapi poin yang bisa kami sampaikan adalah dengan adanya pengembangan library pada python dapat mempermudah pengguna untuk melakukan klasifikasi data dengan cepat dan akurat.

UCAPAN TERIMA KASIH

Terima kasih saya berikan kepada pengembang library Lucifer Machine Learning (link: https://github.com/d4rk-lucif3r/LuciferML) yang memberikan sebuah panduan untuk menyelesaikan analisis sederhana pada data kualitas air untuk dikonsumsi atau tidak dapat dikonsumsi.

REFERENSI

1. Rochmi, MN. "Akses air bersih masih jauh banget dari target". Diakses dari: *https://beritagar.id/artikel/editorial/hapuskan-perda- penyebab-ekonomi-biaya-tinggi.* 2016.
2. UNESCO. "Global Climate Change". Diakses dari: *www.unesco.org.* 19 Agustus 2021.
3. 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.
4. Farooqui, A. "Invertigation of a community outbreak of typhoid fever associated with drinking water". BMC Public Health. 2009. 9: 476.
5. Abrischamchi, A., Ebrahimian, A., Tajrishi, M., dan Marino, M. "Case Study: Application of Multicriteria Decision Making to Urban Water Supply". J Water Resour Plann Manage. 2005. 131(4) Halaman 326-335.
6. Cronin, A.A. "Monitoring source and domestic water quality in parallel with sanitary risk identification in Northern Mozambique to prioritise protection interventions". J Water Health. 2006. Volume 4 Halaman 333-345.
7. Shyamala, R. "Physicochemical Analysis of Borwell Water Samples of Telungu Palayam Area in Coimvatore District, Tamilnadu, India". E-Journal of Chemistry. 2008. Volume 5(4), Halaman 924-929.
8. Momba, M.N.M. "Abundance of pathogenic eschericia coli, Salmonella typhimurium and vibrio cholerae in Nkonkobe drinking water source". J Water Health. 2006. Halaman 289-296.
9. Eschol, J. Is fecal contamination of drinking water after collection associated with household water handling and hygiene practices? A study of urban slum households in Hyderabad, India". Journal of Water and Health. 2009. Halaman 145-154.
10. C. Veesommai and Y.Kiyoki. "Critical Contaminate Detection, Classification of Multiple-water-quality-parameters Values and Real-time Notification by rSPA Processes". Surabaya ; IEEE International Electronics Sysposium, 2015.
11. A. Sarkar dan P.Pandey, "River Water Quality Modelling Using Artificial Neural Network Technique". Aquatic Procedia, Vol.4, 2015, Pages 1070-1077.

12. Y.R.Ding, Y.J.Cai, P.D.Sun and B.Chen. "The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality" Journal of Applied Research and Technology. Vol.12, Issue 3, June 2014, Pages 493-499.
13. Noori, Roohollah, Zhiqiang Deng, Amin Kiaghadi, and Fatemeh Torabi Kachoosangi. "How Reliable Are ANN, ANFIS, and SVM Techniques for Predicting Longitudinal Dispersion Coefficient in Natural Rivers?." Journal of Hydraulic Engineering , 2015: 04015039.
14. Shuxiu Liang, Songlin Han, Zhaochen Sun, "Parameter optimization method for the water quality dinamic model based on data-driven theory". China, 2011.
15. Yue Liao, Jianyu Xu, Wenjing Wang , "A method of Water Quality Assessment Based on Biomonitoring and Multiclass *Support Vector Machine*". International Conferences on ESIAT, 2011.
16. Dark-Lucifer. "LuciferML a semi-supervised machine learning Library by dark-lucifer". Diakses pada tanggal: 18 Agustus 2021. Diakses dari: https://github.com/d4rk-lucif3r/*LuciferML*