



VOLUME 02, No 01, Juni 2023

e-ISSN: 2987-206X

<https://ejournal.unib.ac.id/diophantine>,

Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning

Lidya Savitri ^{1*}, Rahmat Nursalim ¹

¹Department of Mathematics, Universitas Bengkulu, Indonesia

* Corresponding Author: lidyasavitri013@gmail.com

Article Information

Article History:

Submitted: 06 15 2023

Accepted: 06 28 2023

Published: 06 30 2023

Key Words:

klasifikasi Air minum, machine learning, algoritma, Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, K-Nearest Neighbor(KNN), XGBoost Classifier

DOI:

<https://doi.org/10.33369/diophantine.v2i01.28260>

Abstract

The need for the provision and service of clean water from time to time is increasing which is sometimes not matched by the ability and knowledge of clean water. The majority of people still do not know whether water is suitable for consumption or not. The quality of drinking water can be distinguished based on the mineral parameters contained in the water. This article will explain the classification of water sample data by applying a Machine Learning Algorithm, which includes modeling with Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, K-Nearest Neighbor(KNN), XGBoost Classifier. Classification models produce varying degrees of accuracy. The highest accuracy is obtained in the Random Forest Classifier model with an accuracy rate of 78%. Analysis of drinking water quality with machine learning algorithms is very easy to understand, because the results of this study produce very simple results so that they are easy to understand.

1. PENDAHULUAN

Air merupakan kebutuhan penting untuk kelangsungan hidup makhluk hidup yang ada di bumi[1]. Di Perkotaan air merupakan kebutuhan yang sangat penting, terutama kebutuhan akan air bersih. Kebutuhan yang semakin tahun semakin meningkat tetapi berbanding terbalik dengan persediaan air bersih yang semakin terbatas yang diakibatkan banyaknya pembangunan yang tidak memperhatikan daerah resapan air yang semakin sempit[2]. Krisis air bersih melanda berbagai negara di dunia, bahkan air bersih yang dapat dikonsumsi oleh manusia hanya sebesar 1% dari total air yang ada. Jumlah air bersih yang kecil menyebabkan susah penduduk mengakses air bersih. berdasarkan data dari WHO sebanyak 663 juta penduduk kesusahan untuk mengakses air bersih [3].

Berdasarkan Peraturan Menteri Kesehatan Nomor 492/MENKES/PER/IV/2010, Pasal (1) ayat (1) Persyaratan Air minum adalah air yang melalui proses pengolahan atau tanpa proses pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum[4]. Untuk memenuhi standar kualitas air minum upaya pengawasan di daerah sumber air sangatlah penting. Dengan adanya pengawasan sumber air akan terjaga dan mampu menghasilkan kualitas air dengan standar yang layak dikonsumsi oleh manusia[5].

Penelitian tentang klasifikasi air minum sudah banyak dilakukan oleh peneliti, khususnya mengenai klasifikasi kualitas air minum menggunakan penerapan machine learning. Penelitian yang dilakukan oleh Aldi dkk, menggunakan metode Naive bayes, Decision Tree, dan K-Nearest Neighbours pada penelitian ini untuk mengetahui tingkat keakuratan yang paling tinggi dan didapat keakuratan yang paling tinggi adalah metode Decision Tree[7]. Sedangkan penelitian yang dilakukan Prismahardi dkk, menggunakan metode

Support Vector Machine, Decision Tree, Naïve Bayes, dan Artificial Neural Network dari penelitian ini didapat tingkat keakuratan yang paling tinggi metode Random Forest Classifier[8].

Pengukuran kualitas air minum menggunakan parameter dan variabel, dengan pengambilan dataset pada kaggle yang berjudul *water_potability.csv* yang memuat sepuluh parameter. Pada penelitian ini penulis melakukan klasifikasi kualitas air minum dengan menggunakan penerapan Machine Learning, yang mencakup pemodelan dengan Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, K-Nearest Neighbor(KNN), XGBoost Clasifier. Dalam penelitian ini menggunakan pendekatan supervised learning menggunakan library phyton dalam mengelola data. Metode ini digunakan untuk menganalisis data dalam melakukan prediksi dan klasifikasi.

2. METODE

Penelitian ini merupakan penelitian dengan menggunakan data yang diambil dari Kaggle dengan nama *water potability* yang memiliki format csv diselesaikan dengan 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

Potability adalah indikator air yang layak konsumsi dan tidak konsumsi.

Pada penelitian ini terdapat beberapa tahapan dalam mengelolah data diantaranya sebagai berikut :

a. Pre-processing data

Pre-pocessing data yang dilakukan disini adalah klassifikasi yang berarti proses memprediksi kelas atau kategori data dengan memanfaatkan nilai yang ada pada data. Algoritma *machine learning* bagi menjadi dua, yaitu *supervised* dan *unsupervised learning*. Penelitian ini menggunakan metode supervised learning. Pada tahapan awal pengolahan data ini adalah melakukan *Exploratory Data Analysis (EDA)* pada tahapan ini

berfungsi untuk menganalisis data, pada penelitian ini pada EDA menganalisis tipe data, *missing value*, serta menganalisis banyaknya data yang layak dikonsumsi dan tidak untuk dikonsumsi

b. Pemodelan

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 kemungkina[9].

Support Vector Machine (SVM)

SVM merupakan algoritma klasifikasi yang menggunakan ruang hipotesis berupa fungsi-fungsi linear dalam sebuah ruang berdimensi fitur tinggi. SVM memiliki tujuan menemukan fungsi pemisah terbaik antara kelas. Fungsi pemisah paling baik apabila fungsi pemisah yang menghasilkan nilai margin paling besar antara dua vektor dari dua kelas yang berbeda dan berada ditengah-tengah kedua vektor tersebut. Dalam penelitian ini, fungsi pemisah yang dicari adalah fungsi linear[10].

Random Forest Classifier

Random forest adalah algoritma klasifikasi dan regresi yang menjadi bagian dari kelompok *ensemble learning*. Metode *random forest* merupakan pengembangan dari *decision tree* dimana setiap *decision tree* telah dilakukan proses pelatihan dengan menggunakan sampel individu. *Random forest* yang dihasilkan memiliki banyak *tree* dan setiap *tree* ditanam dengan cara yang sama. Seiring dengan bertambahnya dataset, maka *tree* juga ikut berkembang[11].

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah al-goritma yang digunakan untuk melakukan klasifikasi terhadap suatu objek, berdasarkan k buah data latih yang jaraknya paling dekat dengan objek tersebut. Syarat nilai k adalah tidak boleh lebih besar dari jumlah data latih, dan nilai k harus ganjil dan lebih dari satu. Dekat atau jauhnya jarak data latih yang paling dekat dengan objek yang akan diklasifikasi dapat dihitung dengan menggunakan metode *cosine similiarity*[12].

XGBoost Classifier

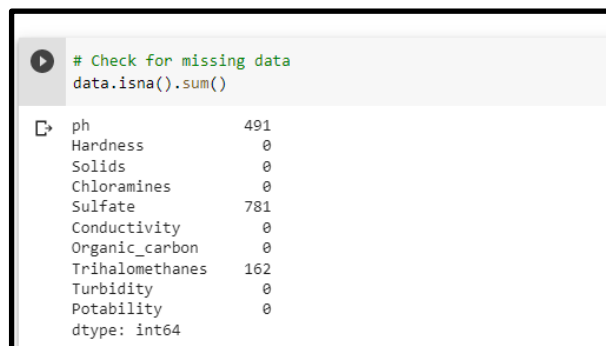
XGBoost Classifier adalah implementasi dari pohon keputusan yang didorong gradien yang dirancang untuk kecepatan dan kinerja. XGBoost adalah algoritma peningkatan gradien ekstrim. Dan itu berarti ini adalah algoritma pembelajaran mesin yang besar dengan banyak bagian. XGBoost bekerja dengan kumpulan data yang besar dan rumit. XGBoost adalah teknik pemodelan ensemble[13].

HASIL DAN PEMBAHASAN

Penelitian ini dimulai dengan mengidentifikasi dataset yang digunakan adalah data dalam format csv yang diperoleh dari kaggle yang berjudul water_potability. Selanjutnya diperiksa tipe data yang digunakan pada dataset. Berikut ini adalah gambar yang menunjukkan tipe dataset yang digunakan. Berdasarkan hasil pemeriksaan data dapat diketahui bahwa ada 2 tipe data yaitu float64 dan int64. Hal ini menunjukkan bahwa semua data bertipe numeric, yang dapat dilihat pada tabel berikut ini.

Data	Parameter	Non-null count	Type
0	pH	2785 non-null	float64
1	Hardness	3276 non-null	float64
2	Solids	3276 non-null	float64
3	Chloramines	3276 non-null	float64
4	Sulfate	2495 non-null	float64
5	Conductivity	3276 non-null	float64
6	Organic Carbon	3276 non-null	float64
7	Trihalomethanes	3114 non-null	float64
8	Turbidity	3276 non-null	float64
9	Potability	3276 non-null	Int64

Langkah selanjutnya dalam mendeskripsi data adalah mengidentifikasi missing value dengan tujuan untuk mengetahui jumlah error dalam data. Berikut ini adalah gambar yang menunjukkan missing value pada dataset. Bisa dilihat bahwa terdapat missing value pada pH, Sulfate, dan trihalomethanes dan masing-masing memiliki missing value sebanyak 491, 781, dan 162. Bisa dilihat bahwa terdapat missing value pada pH, Sulfate, dan trihalomethanes dan masing-masing memiliki missing value sebanyak 491, 781, dan 162.



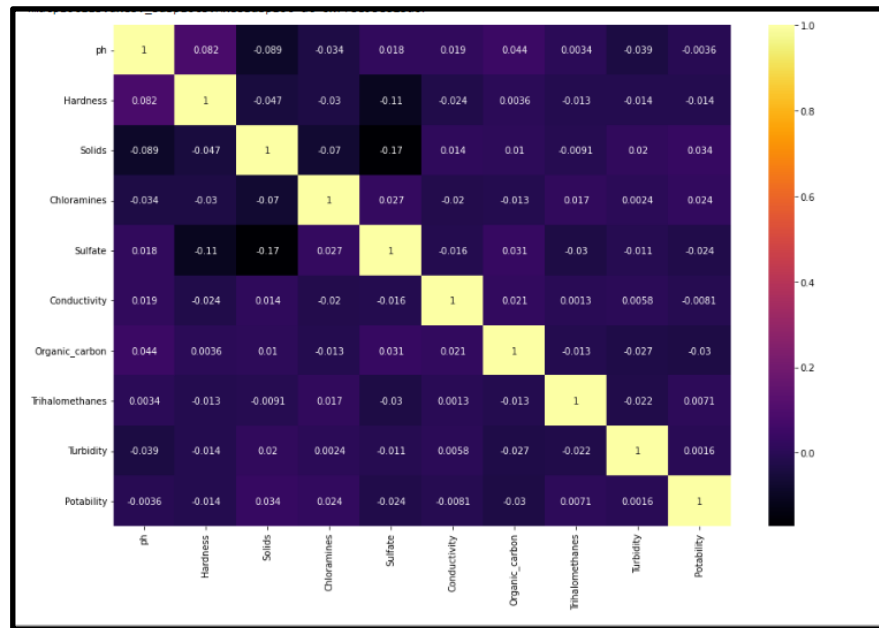
```
# Check for missing data
data.isna().sum()
```

ph	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic_carbon	0
Trihalomethanes	162
Turbidity	0
Potability	0

dtype: int64

Gambar 1 missing value

Langkah selanjutnya adalah menemukan hubungan antar parameter dengan parameter lainnya didalam data. Berikut ini, gambar yang menunjukkan hubungan antar parameter (*correlation matrix*).



Gambar 2 Correlation matrix

Dari Gambar 3 terdapat tiga parameter yang memiliki hubungan antar parameter yang rendah diantaranya *Sulfate-Solids* sebesar 0,17 poin dan *Sulfate-Hardness* 0,11 poin. Selain itu nilai korelasi yang dihasilkan terlalu rendah maka tidak dapat mengidentifikasi hubungan yang jelas antar variabel. Dari indikator warna dapat kita baca, bahwa semakin cerah warna yang tertera korelasi antara dua parameter tersebut semakin besar, begitu sebaliknya apabila indikator warna semakin pekat, korelasi antara dua parameter tersebut semakin kecil.

Langkah selanjutnya Pada library phyton apabila terdapat missing value kita dapat mengatasi *missing value*, salah satu nya dengan menggunakan fungsi *def fill nan*, dimana nilai NaN akan diganti dengan nilai mean. Setelah melakukan proses ini dapat kita lihat pada Gambar 4 bahwa tidak ada lagi missing value pada data.

```
def fill_nan(df):
    for index, column in enumerate(df.columns[:9]):
        # print(index, column)
        df[column] = df[column].fillna(df.groupby('Potability')[column].transform('mean'))
    return df

df = fill_nan(data)

df.isna().sum()

ph          0
Hardness    0
Solids       0
Chloramines  0
Sulfate      0
Conductivity 0
Organic_carbon 0
Trihalomethanes 0
Turbidity    0
Potability   0
dtype: int64
```

Gambar 3 def fill nan

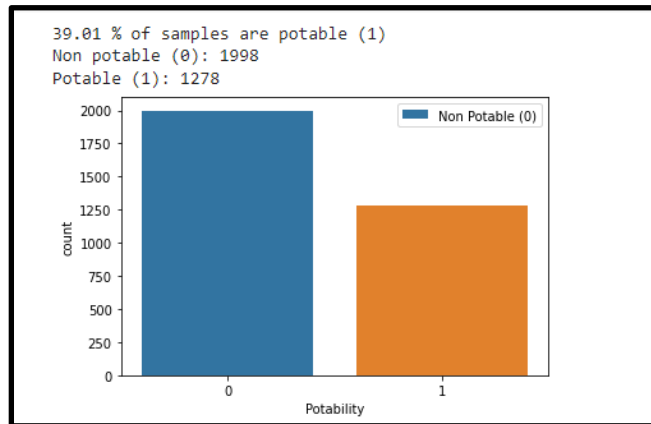
Dimana missing value tersebut diganti dengan nilai mean dari data, dimana nilai mean dari disajikan pada gambar 3.5 berikut ini.

	count	mean	std	min	25%	50%	75%	max
ph	3276.0	7.080855	1.469958	0.000000	6.277673	7.085378	7.870050	14.000000
Hardness	3276.0	196.369496	32.879761	47.432000	176.850538	196.967627	216.667456	323.124000
Solids	3276.0	22014.092526	8768.570828	320.942611	15666.690297	20927.833607	27332.762127	61227.196008
Chloramines	3276.0	7.122277	1.583085	0.352000	6.127421	7.130299	8.114887	13.127000
Sulfate	3276.0	333.785123	36.145701	129.000000	317.094638	334.564290	350.385756	481.030642
Conductivity	3276.0	426.205111	80.824064	181.483754	365.734414	421.884968	481.792304	753.342620
Organic_carbon	3276.0	14.284970	3.308162	2.200000	12.065801	14.218338	16.557652	28.300000
Trihalomethanes	3276.0	66.395671	15.769901	0.738000	56.647656	66.303555	76.666609	124.000000
Turbidity	3276.0	3.966786	0.780382	1.450000	3.439711	3.955028	4.500320	6.739000
Potability	3276.0	0.390110	0.487849	0.000000	0.000000	0.000000	1.000000	1.000000

Gambar 5 analysis statically

Nilai Mean diatas yang menggantikan nilai NaN pada pH, Sulfate, dan Trihalomethanes yang sebelumnya hilang.

Langkah selanjutnya adalah mengkategorikan kualitas air yang diwakilkan oleh dataset yang disajikan dengan Exploratory Data Analysis (EDA). Berikut gambar yang menunjukkan kategori kualitas air yang potability dan non potability



Gambar 6 Diagram kualitas air minum berdasarkan database

Dari diagram diatas diketahui dari total 3276 data terdapat 1998 data non potable (tidak layak dikonsumsi) dan 1278 data potable (layak dikonsumsi). Dapat dibuat dalam persen bahwa dari total dari terdapat 60.99% non potable dan 31.01% potable.

Langkah selanjutnya menjalankan algoritma machine learning. Semua algoritma dituliskan dalam bahasa pemrograman python yang dijalankan secara bersama.

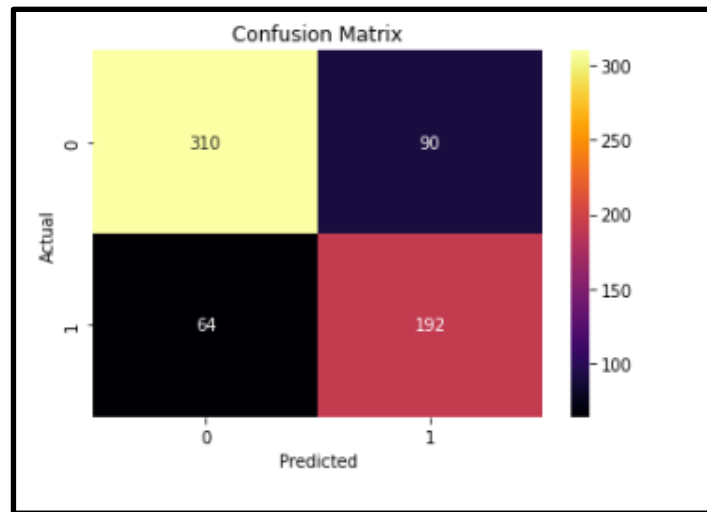
Proses klasifikasi kualitas air minum melewati proses modelling yang dilakukan dengan menggunakan lima algoritma machine learning menghasilkan nilai yang berbeda antara algoritma satu dengan yang lain. Tabel 1 akan menunjukkan hasil dari proses modelling lima algoritma tersebut

Tabel 1 Hasil Klasifikasi

Classifier	Class	Precision	Recall	f1-score	Akurasi
Logistic Regression	0	0.62	0.50	0.56	0.51
	1	0.40	0.53	0.46	
SVM	0	0.73	0.64	0.68	0.64
	1	0.53	0.64	0.58	
XGB	0	0.85	0.72	0.78	0.75

Classifier	1	0.65	0.80	0.72	
Random Forest	0	0.83	0.77	0.80	0.77
	1	0.68	0.76	0.72	
KNN	0	0.72	0.62	0.67	0.62
	1	0.51	0.62	0.56	

Berdasarkan hasil diatas dapat disimpulkan bahwa algoritma random forest dan XGBclassifier memiliki nilai yang paling besar. namun, disini kita kan menggunakan algoritma Random Forest untuk kita olah pada final model dengan confusion matrix, hal ini dikarenakan dalam pemilihan algoritma itu harus dipilih algoritma yang memiliki akurasi, precision, recall serta f1-score, maka kita akan mengambil random forest memenuhi tiga persyaratan dari empat persyaratan.



Gambar 7 Confusion Matrix Random Forest Classifier

Berdasarkan Gambar 7 akan diketahui nilai pesisi, recall, dan akurasi pada random forest classifier. Nilai yang dihasilkan gambar diatas dapat diartika bahwa nilai 310 yang berarti True Positive (TP), nilai 192 yang berarti True Negative (TN), nilai 90 berarti False Positif (FP), dan nilai 64 berarti False Negative (FN).

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

$$\text{Presisi} = \text{TP}/(\text{TP}+\text{FP})=310/(310 + 90)=310/400=0.775=77.5\%$$

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

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})=310/(310 +64)=310/374=0.8288=82.88\%$$

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

$$\begin{aligned} \text{Akurasi} &= (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{TN}+\text{FN})=(310 + 192)/(310+ 90+192+64)= 502/656 \\ &=0.7652=0.77=77\% \end{aligned}$$

Sehingga diperoleh nilai presisi 77.5%, recall 82.88%, dan akurasi 76%..

3. SIMPULAN

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 beberapa algoritma lainnya yang digunakan. Sistem identifikasi kualitas air minum memiliki akurasi sebesar 76.52% dengan menggunakan algoritma Random Forest Classifier.

REFERENSI

- [1] Hamidi, Rifwan, M. Tanzil Furqon, and Bayu Rahayudi. "Implementasi Learning Vector Quantization (LVQ) untuk Klasifikasi Kualitas Air Sungai." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN 2548* (2017): 964X.
- [2] Arnomo, Rio Adi. *Implementasi Algoritma K-Nearest Neighbor untuk Identifikasi Kualitas Air (Studi Kasus: PDAM Kota Surakarta)*. Diss. STMIK Sinar Nusantara Surakarta, 2017.
- [3] Riyantoko, Prismahardi Aji, Tresna Maulana Fahrudin, and Kartika Maulida Hindrayani. "Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning." *SENADA 1.01* (2021): 12-18.
- [4] Indonesia, Pemerintah Republik. "Peraturan Menteri Kesehatan Nomor 492/Menkes/Per/IV/2010 tentang Baku Mutu Air Minum." (2010).
- [5] Tumangger, Ricky Marten Sahalutua. *Komparasi Metode Data Mining Support Vector Machine Dengan Naive Bayes Untuk Klasifikasi Status Kualitas Air*. Diss. Universitas Brawijaya, 2020.
- [6] Tangkelayuk, Aldi. "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree." *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)* 9.2 (2022): 1109-1119.
- [7] Bowo Djoko Marsono, Yoga Ardy Pradana dan. "Uji Kualitas Air Minum Isi Ulang di Kecamatan Sukodono, Sidoarjo Ditinjau dari Perilaku dan Pemeliharaan Alat." *JURNAL TEKNIK POMITS*, vol. vol.2, pp. D-83
- [8] Vidiastanta, Icha Gusti, Nurul Hidayat, and Ratih Kartika Dewi. "Komparasi Metode K-Nearest Neighbors (K-NN) Dengan Support Vector Machine (SVM) Untuk Klasifikasi Status Kualitas Air." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN 2548* (2020): 964X.
- [9] S. Majumder, A. Aich, and S. Das, "Sentiment Analysis of People During Lockdown Period of COVID-19 Using SVM and Logistic Regression Analysis," *SSRN Electron. J.*, Mar. 2021, doi: 10.2139/SSRN.3801039.
- [10] Perdana, A. dan Furqon, M. T. (2018) "Penerapan Algoritma Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia (Studi Kasus : RSJ . Radjiman Wediodiningrat , Lawang)," 2(9), hal. 3162–3167.
- [11] Saleh, A. (2015) "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," 2(3), hal. 207–217.
- [12] J. Han, M. Kamber dan J. Pei, *Data Mining Concepts and Techniques*, Waltham: Elsevier, 2012.
- [13] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).