Pengembangan Sistem Klasifikasi Kualitas Air Minum Berbasis Web Menggunakan Algoritma *K-Nearest Neighbors*

1st Ivana Meiska
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesa
ivanameiska@student.telkomuniversity.ac.id

2th Meta Kallista
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesa
metakallista@telkomuniversity.ac.id

3th Ig.Prasetya Dwi Wibawa
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesa
prasdwibawa@telkomuniversity.ac.id

***Abstrak* — Air memiliki peran penting sebagai kebutuhan primer dalam kehidupan manusia, termasuk untuk konsumsi. Namun, sayangnya air mudah terkontaminasi sehingga dapat membahayakan kesehatan tubuh. Oleh karena itu, penentuan kelayakan air minum dengan metode manual seperti STORET dan Indeks Pencemaran memakan waktu lama dan biaya yang tinggi. Untuk mengatasi hal ini, penerapan machine learning dengan algoritma *K-Nearest Neighbors* dan teknik SMOTE untuk mengatasi ketidakseimbangan pada kelas target menjadi pilihan yang efisien dan tepat. Hasil penelitian menunjukkan bahwa model *K-Nearest Neighbors* dengan k=3 mampu mencapai akurasi training sebesar 0.98928, akurasi testing sebesar 0.99434, serta ROC AUC mencapai 1.00 dengan *loss* hanya 0.38618. Model yang optimal akan divisualisasikan hasilnya menggunakan Streamlit ssebagai alat untuk menyajikan informasi secara interaktif, memungkinkan pengguna untuk dengan mudah memahami dan menganalisis kualitas air minum.**

Kata kunci—Kelayakan Air Minum, K-Nearest Neighbors, Machine Learning, SMOTE, Streamlit.

I. PENDAHULUAN

Air merupakan salah satu kebutuhan primer bagi manusia dalam kehidupan sehari-hari, seperti mencuci, mandi dan dikonsumsi. Sumber air dapat berasal dari mana saja, seperi hujan, sungai, air tanah dan lainnya. Di Indonesia, air yang layak konsumsi diatur dalam Peraturan Menteri Kesehatan Republik Indonesia Nomor 2 Tahun 2023. Peraturan ini penting, mengingat bahwa air mudah terkontaminasi dan dikhawatirkan apabila air tersebut dikonsumsi dapat membahayakan kesehatan tubuh. Oleh sebab itu, penting untuk mengetahui kelayakan atau kualitas air yang akan dikonsumsi[1–3].

Untuk mengetahui kelayakan atau kualitas air, perlu dilakukan pengujian. Ada beberapa metode manual yang umum digunakan mengklasifikasi kelayakan air minum berdasarkan kandungannya, yaitu STORET (*STOrage and RETrieval*) dan Indeks Pencemaran. Kedua metode ini memiliki kelemahannya masing-masing sehingga menyebabkan kurangnya efisiensi dari segi waktu, tenaga manusia, biaya dan juga kurangnya akurasi dari hasil metode ini[4].

Berdasarkan kelemahan-kelemahan metode di atas, peneliti mengusulkan agar menggunakan *machine learning* untuk mengklasifikasi kelayakan air minum berdasarkan kandungan parameter airnya. Pengembangan yang tepat untuk menentukan algoritma dan teknik yang paling cocok dapat dicapai melalui eksperimen dan perbandingan antarmetode. Salah satu penelitian yang dilakukan pada tahun 2022 oleh Sai Sreeja dan teman-teman meneliti perbandingan algoritma *K-Nearest Neighbors* dengan algoritma *Decision Tree* untuk mengklasifikasi kualitas air dengan 9 *features*. Hasil yang didapatkan dari penelitian ini yaitu algoritma *K-Nearest Neighbors* memiliki akurasi yang lebih tinggi sebesar 61.7% dibandingkan algoritma *Decision Tree* yang memiliki akurasi rata-rata sebesar 58.5%[5].

Pada tahun 2020, penelitian yang dilakukan oleh A G Pertiwi dan teman-teman mengenai perbandingan performa algoritma *K-Nearest Neighbors* saat kondisi *imbalanced data* dengan data yang sudah diseimbangkan menggunakan SMOTE dalam studi kasus diagnosis penyakit diabetes. Dari penelitian ini didapatkan kesimpulan bahwa akurasi yang dihasilkan dalam mendiagnosa penyakit diabetes menggunakan *K-Nearest Neigbors* dengan SMOTE lebih baik dibandingkan dengan akurasi yang dihasilkan menggunakan *K-Nearest Neighbors* tanpa SMOTE untuk kasus dataset yang tidak seimbang[6].

Berdasarkan kedua permasalahan dan penelitian yang dipaparkan di atas, peneliti mengusulkan agar membuat suatu sistem menggunakan *machine learning* yang datasetnya sudah diseimbangkan menggunakan teknik SMOTE. *Machine learning* digunakan untuk mengklasifikasi menggunakan algoritma *K-Nearest Neighbors,* lalu hasil klasifikasi akan divisualisasikan menggunakan *website*. Penjelasan mengenai penelitian ini akan dipaparkan pada Bagian II.

II. KAJIAN TEORI

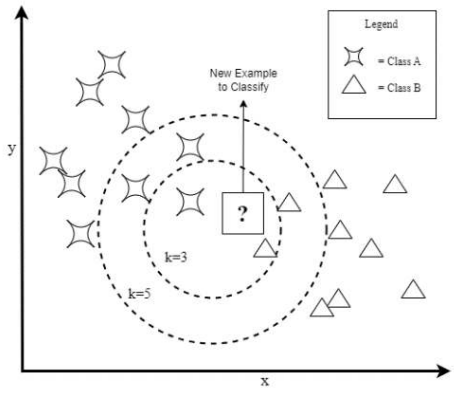
A. *Machine Learning*

Machine learning atau pembelajaran mesin merupakan suatu proses yang memungkinkan mesin atau komputer dapat menganalisis dan mempelajari pola secara mandiri berdasarkan data yang telah diberikan agar dapat memberikan prediksi dan mengambil keputusan. Semakin banyak data pada pola yang diajarkan, maka mesin atau komputer akan semakin handal dalam melakukan tugasnya[7].

B. Klasifikasi dan Algoritma *K-Nearest Neighbors*

Klasifikasi adalah suatu proses pengelompokkan benda berdasarkan karakteristik yang dimiliki oleh objek klasifikasi. Klasifikasi dapat dilakukan secara manual ataupun dengan bantuan teknologi, salah satunya *machine learning*[8]. Ada beberapa algoritma yang dapat melakukan klasifikasi seperti *K-Nearest Neighbors*, *Decision Tree*, *K-Means* dan lainnya. Pada penelitian ini hanya berfokus pada algoritma *K-Nearest Neighbors*.

Konsep kerja dari algoritma ini yaitu berbasis pada jarak (*Instanced Base Learning*), yang mana algoritma ini akan mencari jarak terdekat dengan *k* tetangganya pada *data training* dengan data yang akan diuji[9]. Nilai *k* pada algoritma *K-Nearest Neighbors* sangat penting, karena berpengaruh pada performa klasifikasi yang dihasilkan[10]. Ilustrasi dari *K-Nearest Neighbors* dapat dilihat pada GAMBAR 1.



GAMBAR 1.
Ilustrasi algoritma K-NN

Ada beberapa fungsi yang digunakan untuk menghitung jarak dalam *K-Nearest Neighbors*, tetapi pada umumnya digunakan fungsi *Euclidean Disctance*, dengan rumus:

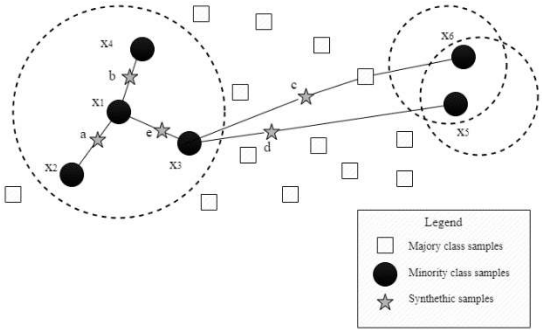
$$euc = \sqrt{\sum_{i=1}^{n}(x_{2i} - x_{1i})^2} \quad (1)$$

Keterangan :

- x_2 = *data training*
- x_1 = *data testing*
- i = variable data
- n = dimensi data

C. *Imbalanced Data* dan SMOTE

Pada *machine learning* kerap terjadi kondisi dimana kelas pada target di dataset latih tidaklah seimbang[11]. Contoh dari *imbalanced data* atau ketidakseimbangan data yaitu, ketika sebuah dataset memiliki target kelas 'layak' sebanyak 2345 data sementara target kelas 'tidak layak' sebanyak 123 data. Dapat dilihat bahwa adanya salah satu kelas yang mendominasi suatu target, yang menyebabkan kelas lainnya menjadi minor. Kondisi ini mengakibatkan model *machine learning* yang dibangun cenderung memprediksi kelas mayoritas dengan akurasi yang tinggi, sementara saat memprediksi kelas minoritas tidak jarang data diabaikan atau dianggap sebagai *noise*[12]. Untuk menangani masalah ketidakseimbangan ini perlu dilakukan penyeimbangan data. Ada 2 teknik penanganan *imbalanced data* yaitu *oversampling* atau *undersampling*. Pada penelitian ini, digunakan teknik *oversampling* menggunakan metode SMOTE yang divisualisasikan pada GAMBAR 2.



GAMBAR 2.
Ilustrasi SMOTE

Synthetic Minority Oversampling Technique (SMOTE) merupakan salah satu teknik penanganan masalah *imbalanced data* dengan konsep kerja membuat replikasi atau menyintesiskan data minoritas. Konsep kerja teknik SMOTE sama dengan algoritma *K-Nearest Neighbors*, yaitu mencari ketetanggan terdekat sebanyak nilai *k* untuk setiap data di kelas minoritas. Setelah didapat ketetanggan terdekatnya, data akan direplikasi atau disintesis sebanyak persentasi duplikasi yang diinginkan antara data minor dan nilai *k* yang dipilih secara acak[10]

D. Evaluasi Model

Setelah membangun model *machine learning*, dilakukan pengevaluasian model untuk memvalidasi hasil pengujian yang telah dilakukan pada model *machine learning* yang telah dibangun. Evaluasi model yang akan digunakan pada penelitian ini yaitu akurasi, *confusion matrix* dan kurva AUC-ROC.

Confusion matrix adalah tabel yang merepresentasikan kinerja model klasifikasi dengan konsep kerja yaitu membandingkan prediksi model dengan nilai yang sebenarnya yang diuji. Nilai aktual ditandai dengan *True* (1) dan *False* (0), lalu diprediksi sebagai *Positive* (1) dan *Negative* (0)[13]. *Confusion matrix* ditampilkan pada GAMBAR 3.

Class Designation		Actual Class	
		True (1)	False (0)
Predicted Class	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

GAMBAR 3.
Confusion Matrix

Dari *confusion matrix*, dapat ditentukan nilai akurasi, precision, *recall* dan *F1-Score*. Akurasi merupakan penjumlahan pengujian yang benar antara sampel yang diklasifikasikan dengan jumlah total. *Precision* merepresentasikan nilai positif yang diklasifikan benar ke dalam jumlah total sampel yang diprediksi positif. *Recall* adalah adalah kemampuan model untuk mengidentifikasi dan mendeteksi contoh positif secara benar dari keseluruhan contoh positif yang sebenarnya[14]. *F1-Score* adalah matrik yang kerap digunakan untuk mengevaluasi masalah *imbalanced class*. Konsep evaluasi dari *F1-Score* yaitu menggabungkan *recall* dan *precision*, sehingga akan

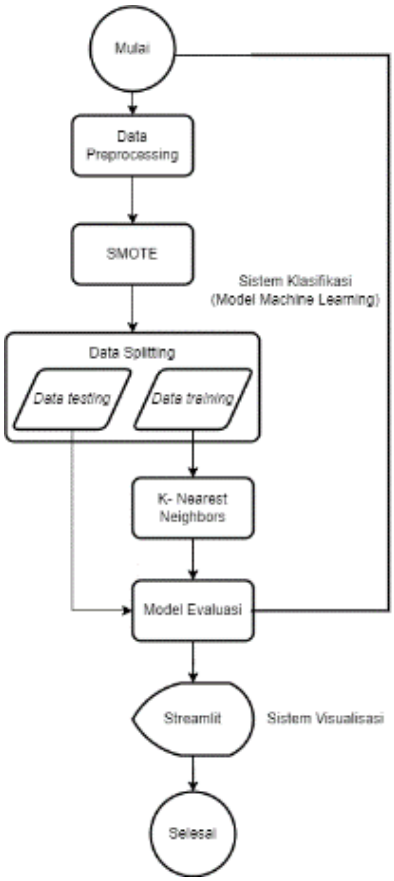
menghasilkan metrik yang efektif untuk mencari kembali informasi dalam dataset yang mengandung ketidakseimbangan data[15]. Kurva ROC adalah grafik evaluasi yang memvisualisasikan *True Positive* dan *True False*. Semakin tinggi *True Positive* maka semakin rendah *True False*. Area yang berada di bawah kurva ROC disebut AUC merepresentasikan seberapa baik model *machine learning* dalam memprediksi. Semakin tinggi nilai AUC, semakin baik model yang dibangun[13].

E. Streamlit

Streamlit merupakan *framework open source* berbasis Python yang dirancang untuk memudahkan pengembang dalam membuat aplikasi web interaktif di bidang *data science* dan *machine learning*. Salah satu keunggulan Streamlit adalah pengembang tidak perlu mengurus tampilan *website* menggunakan CSS, HTML, dan JavaScript secara manual, karena *framework* Streamlit menyediakan fungsi-fungsi yang sudah siap digunakan untuk mengatur tampilan tersebut[16].

III. METODE

Penelitian ini terdiri dari 2 sistem, yaitu Sistem Klasifikasi dan Sistem Visualisasi. Alur rencana penelitian ini ditampilkan pada GAMBAR 4.



GAMBAR 4.
Alur penelitian

A. Sistem Klasifikasi

Dataset yang digunakan pada penelitian ini berasal dari berbagai sumber, salah satunya berasal dari PDAM dan perusahaan air minum. *Dataset* ini terdiri dari 27 *features* sebagai parameter masukan dan 1 target dengan 2 kelas yaitu 1 sebagai layak dan 0 sebagai tidak layak. Parameter ini dapat dilihat pada TABEL 1.

TABEL 1.
Parameter *features*

Parameter	Batas Max	Parameter	Batas Maksimal
E.Colli	0 jml/100 mL	Sianida	0.07 mg/L
Coliform	0 jml/100 mL	Selenium	0.01 mg/L
Arsen	0.01 mg/L	Alumunium	0.2 mg/L
Kromium	0.05 mg/L	Besi	0.3 mg/L
Kadmium	0.003 mg/L	Kesadahan	500 mg/L
Nitrit	3 mg/L	Klorida	250 mg/L
Nitrat	50 mg/L	Mangan	0.4 mg/L
pH	6.5 - 8.5	Bau	Tidak Berbau : 1
Seng	3 mg/L	Warna	15 TCU
Sulfat	250	TDS	500 mg/L
Tembaga	2 mg/L	Kekeruhan	5 NTU
Amonia	1.5 mg/L	Rasa	Tidak Berasa : 1
Chlor	0.2 - 1.0 mg/L	Suhu	Suhu +- 3 C
BOD5	- mg/L	COD	- mg/L

Pada proses *data preprocessing*, dilakukan *data cleaning* untuk mengatasi *missing value*, lalu ada proses *data transformation* dengan tujuan mengubah tipedata *object* pada parameter ‘Bau’ menjadi tipedata int64 dan tipedata ‘Rasa’ menjadi int64. Setelah proses *data preprocessing*, dataset dianggap ‘bersih’ sehingga dapat dilakukan penanganan *imbalanced data* pada dua kelas di target menggunakan SMOTE. Proses selanjutnya yaitu *data splitting*. Dalam penelitian ini akan digunakan proporsi 80% *data training* dan 20% *data testing*.

Setelah kelas target pada *dataset* seimbang, langkah selanjutnya yaitu memasukkan algoritma *K-Nearest Neighbors* untuk mengklasifikasi data yang masukkan. Pada tahap ini, peneliti melakukan percobaan dari k=1 hingga k=45 lalu dipilih nilai k terbaik berdasarkan evaluasi model.

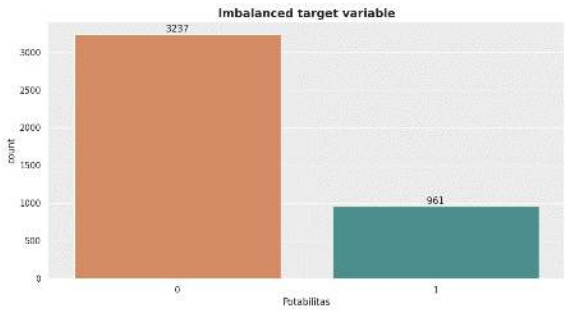
B. Sistem Visualisasi

Setelah mendapat model *machine learning* yang akan digunakan untuk mengklasifikasi, selanjutnya model tersebut disimpan dalam formal .sav pada Github. Setelah disimpan di Github, model *machine learning* dapat diterapkan di Streamlit menggunakan Visual Studio Code.

IV. HASIL DAN PEMBAHASAN

Pada bagian ini, peneliti akan memaparkan hasil dan analisis dari penelitian yang sudah dirancang. Penelitian ini dilakukan menggunakan komputer dengan spesifikasi CPU Intel Core i7 -8550U 1.80GHz, RAM 8GB, dan sistem operasi Microsoft Windows 10 Home Single Language 64-bit. Platform yang digunakan untuk eksekusi *machine learning* adalah Google Colaboratory.

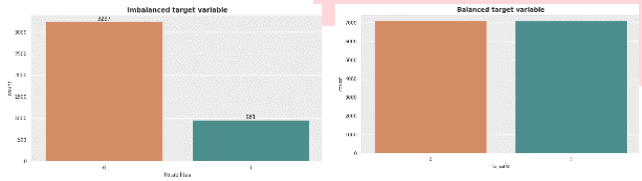
Pada penelitian ini, setelah melakukan *preprocessing data* diketahui bahwa terjadi kondisi *imbalanced data* pada kelas target. Kondisi *imbalanced data* dapat dilihat pada GAMBAR 5.



GAMBAR 5.
Imbalanced data

Dari GAMBAR 5 dapat dilihat bahwa kelas 0 mendominasi dengan total data sebanyak 3237 dibanding dengan total data kelas 1 sebanyak 961. Kondisi ini harus ditangani karena sangat berpengaruh terhadap kinerja model.

Pada penelitian ini, digunakan Teknik SMOTE untuk menangani *imbalanced data*. GAMBAR 6 menunjukkan dua kelas pada target (a) sebelum SMOTE dan (b) sesudah SMOTE.



(a) (b)

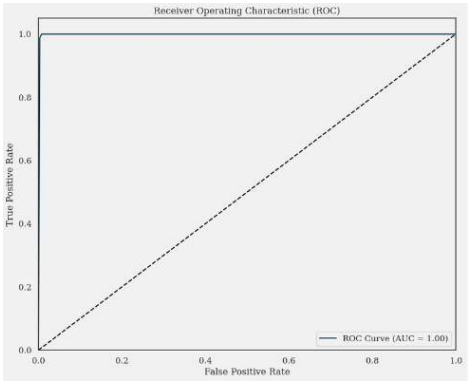
GAMBAR 6
(a) sebelum SMOTE dan (b) setelah SMOTE

Setelah penanganan *imbalanced data*, dilakukan proses klasifikasi menggunakan algoritma *machine learning*. Dari percobaan *k*=1 hingga *k*=45, dipilih 5 percobaan dengan nilai akurasi tertinggi, yaitu seperti TABEL 2.

TABEL 2.
Hasil klasifikasi K-NN setelah SMOTE

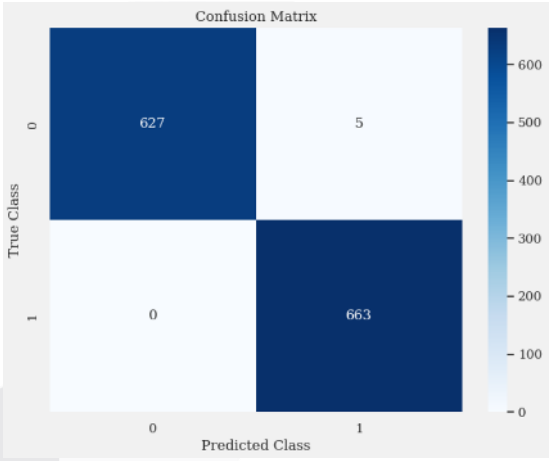
k	Akurasi	Nilai AUC
1	0.996	1.00
2	0.994	1.00
3	0.996	1.00
4	0.995	1.00
5	**0.997**	**1.00**

Dari TABEL 2 dapat dilihat bahwa *k*=5 memiliki akurasi tertinggi sebesar 0.997 dan menghasilkan ROC AUC 1.00, seperti pada GAMBAR 7 di bawah ini. Hal ini menandakan bahwa model dapat membedakan kelas positif dan *negative* secara optimal.



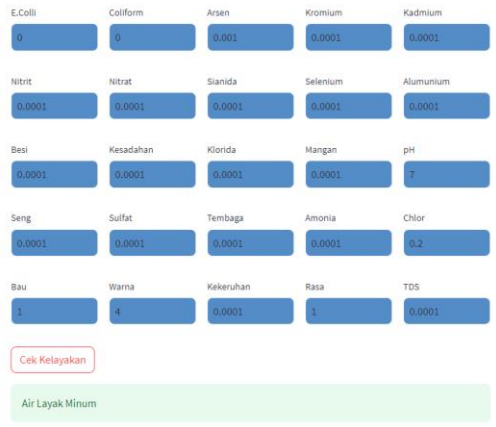
GAMBAR 7.
ROC AUC k=5

Model *machine learning* yang baik ini juga didukung dengan hasil evaluasi menggunakan *confusion matrix* yang divisualisasikan pada GAMBAR 8 di bawah ini. Hasil dari confusion matrix ini menunjukkan bahwa model memiliki performa yang sangat baik dalam mengidentifikasi data positif (sensitivitas yang tinggi) dengan 627 True Positive dan tidak mengalami kesalahan dalam memprediksi data negatif (spesifisitas yang tinggi) dengan 663 True Negative dan 0 False Positive. Namun, ada 5 data positif yang salah diprediksi sebagai negatif (False Negative), yang perlu menjadi perhatian untuk peningkatan performa model di masa mendatang.

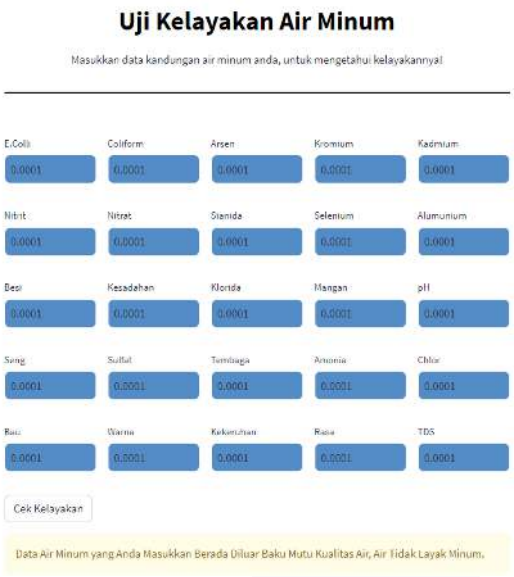


GAMBAR 8.
Confusion matrix k=5

Model *machine learning* terbaik ini akan diimplementasikan ke *website* menggunakan Streamlit. GAMBAR 9 menunjukkan ketika *website* menampilkan klasifikasi air layak minum dan GAMBAR 10 menunjukkan ketika *website* menampilkan klasifikasi air tidak layak minum.



GAMBAR 9.
Tampilan *website* saat mengklasifikasi air layak minum



GAMBAR 10.
Tampilan *website* saat mengklasifiksasi air tidak layak minum

V. KESIMPULAN

Berdasarkan penelitian ini, dapat disimpulkan bahwa teknik SMOTE mampu menyeimbangkan kelas pada target. Algoritma *K-Nearest Neighbors*. Dari penelitian yang telah dicoba dari k=1 hingga k=45 didapat bahwa k=5 memiliki akurasi tertinggi yaitu sebesar 0.997. Oleh sebab itu, dapat dikatakan bahwa percobaan yang dilakukan berulang kali (*tunning* parameter) sangat penting untuk dilakukan agar mendapat model *machine learning* berkali. Selain ini penggunaan Streamlit untuk memvisualisasikan hasil dari model *machine learning* yang telah dibuat cukup efektif dan mudah, mengingat bahwa Streamlit dirancang untuk pengembangan *machine learning.*

REFERENSI

[1] S. Sharma and A. Bhattacharya, "Drinking water contamination and treatment techniques," *Applied Water Science*, vol. 7, no. 3. Springer Verlag, pp. 1043–1067, Jun. 01, 2017. doi: 10.1007/s13201-016-0455-7.

[2] T. Ling, "A Global Study About Water Crisis," 2022.

[3] "PMK-No-492-ttg-Persyaratan-Kualitas-Air-Minum".

[4] N. V. Sidabutar, D. M. Hartono, T. E. B. Soesilo, and R. C. Hutapea, "The quality of raw water for drinking water unit in Jakarta-Indonesia," in *AIP Conference Proceedings*, American Institute of Physics Inc., Mar. 2017. doi: 10.1063/1.4978140.

[5] S. Sreeja Kurra, S. Geethika Naidu, S. Chowdala, S. Chithra Yellanki, and B. Esther Sunanda, "WATER QUALITY PREDICTION USING MACHINE LEARNING." [Online]. Available: www.irjmets.com

[6] A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada, and A. Wibowo, "Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1524/1/012048.

[7] W. Wang and K. Siau, "Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda," *Journal of Database Management*, vol. 30, no. 1, pp. 61–79, 2019, doi: 10.4018/JDM.2019010104.

[8] A. P. Wibawa, M. Guntur, A. Purnama, M. Fathony Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," *Prosiding Seminar Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 1, 2018.

[9] M. A. Rahman, N. Hidayat, and A. A. Supianto, "Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[10] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," 2018.

[11] A. A. Arifiyanti and E. D. Wahyuni, "SMOTE: METODE PENYEIMBANG KELAS PADA KLASIFIKASI DATA MINING", [Online]. Available: https://www.cs.

[12] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf Sci (N Y)*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.

[13] Ž. Vujović, "Classification Model Evaluation Metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.

[14] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.

[15] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, A. Nikmatul Kasanah, U. Pujianto, T. Elektro, F. Teknik, and U. Negeri Malang, "Terakreditasi SINTA Peringkat 2 Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *masa berlaku mulai*, vol. 1, no. 3, pp. 196–201, 2017.

[16] A. Putranto, N. L. Azizah, I. Ratna, I. Astutik, F. Sains, and D. Teknologi, “Sistem Prediksi Penyakit Jantung Berbasis Web Menggunakan Metode SVM dan Framework Streamlit,” 2023. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease