ISSN: 2808-7771 online

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 16-24, 2024

Prediksi Kualitas Air di Jawa Timur Menggunakan Metode CRISP-DM dengan Algoritma K-NN dan Regresi Logistik Multinomial

Kezia Elice Yulianto[1] dan Valencia Melita Christy[2]

[1,2] Program Studi Sistem Informasi, Universitas Ciputra Surabaya
Citraland CBD Boulevard, Surabaya, Indonesia, 60219

Korespondensi: Kezia Elice Yulianto (kyulianto@student.ciputra.ac.id)

Received: 24 Juli 2024 – *Revised*: 31 Agustus 2024 - *Accepted*: 05 Sept 2024 - *Published*: 10 Sept 2024

Abstrak. Air merupakan salah satu sumber daya penting bagi kehidupan manusia. Kualitas air yang buruk dapat menyebabkan berbagai masalah kesehatan, seperti diare dan stunting, yang merupakan penyebab utama kematian pada balita di Indonesia. Penelitian ini bertujuan untuk memprediksi kualitas air di Jawa Timur menggunakan metode pembelajaran mesin K-NN dan Regresi Logistik Multinomial. Data yang digunakan adalah data kualitas air Jawa Timur yang diambil dari situs Kaggle, mencakup variabel temperatur, TDS, BOD, COD, DO, dan kelas. Setelah melalui metode CRISP-DM, hasil evaluasi menunjukkan bahwa model Regresi Logistik Multinomial memiliki akurasi lebih tinggi (0.72) dibandingkan dengan model K-NN (0.67), sehingga dapat disimpulkan bahwa model Regresi Logistik Multinomial dalam penelitian ini lebih cocok digunakan untuk memprediksi kualitas air di Jawa Timur. Berdasarkan hasil analisis, ditemukan bahwa variabel DO dan TDS memiliki pengaruh terbesar terhadap kualitas air. Tingginya kadar DO menunjukkan air yang layak digunakan, sementara tingginya TDS menunjukkan adanya polusi dan limbah yang berbahaya. Hasil penelitian ini diharapkan dapat membantu meningkatkan kualitas air di Jawa Timur, serta mendukung tercapainya Tujuan Pembangunan Berkelanjutan (SDGs) nomor 6 mengenai air bersih dan sanitasi yang berkelanjutan.

Kata kunci: CRISP-DM, K-NN, kualitas air, prediksi, Regresi Logistik Multinomial

Citation Format: Yulianto, K.E., & Christy, V.M. (2024). Prediksi Kualitas Air di Jawa Timur Menggunakan Metode CRISP-DM dengan Algoritma K-NN dan Regresi Logistik Multinomial. *Prosiding SENAM 2024: Seminar Nasional Sistem Informasi & Informatika Universitas Ma Chung*. 4, 16-24. Malang: Ma Chung Press.

PENDAHULUAN

Air merupakan salah satu sumber daya yang sangat penting bagi manusia. Setiap manusia membutuhkan air bersih untuk berbagai aktivitas, mulai dari mandi, memasak, mencuci, dan yang terpenting adalah untuk dikonsumsi. Namun realitanya, tidak semua air dapat dikonsumsi. Salah satu penyebabnya adalah meningkatnya jumlah penduduk setiap tahunnya, khususnya di Jawa Timur. Menurut Badan Pusat Statistik (2023), jumlah penduduk di Jawa Timur mencapai 41.416.407 jiwa setelah mengalami kenaikan dari

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 16-24, 2024
ISSN: 2808-7771 online

angka 41.149.974 jiwa di tahun sebelumnya. Peningkatan jumlah populasi ini dapat mengakibatkan meningkatnya aktivitas industri sehingga terjadi penurunan kualitas air yang akan berdampak signifikan terhadap kesehatan masyarakat. Seperti yang dapat dilihat pada Gambar 1, Indeks Kualitas Air (IKA) di Jawa Timur tidak mengalami perubahan yang signifikan dalam beberapa tahun terakhir.

Gambar 1. Indeks Kualitas Air (IKA) di Jawa Timur. Sumber: BPS Jawa Timur (2022)

Buruknya kualitas air dapat menyebabkan berbagai masalah kesehatan, seperti diare dan stunting, yang termasuk faktor penyebab utama kematian pada balita berdasarkan catatan Kementerian Kesehatan Indonesia. Hal ini menunjukkan bahwa kualitas air merupakan salah satu hal yang tidak dapat diabaikan. Pentingnya kualitas air ini diakui dalam Tujuan Pembangunan Berkelanjutan atau *Sustainable Development Goals* (SDGs) nomor 6 yang menargetkan terjaminnya ketersediaan dan pengelolaan air bersih serta sanitasi yang berkelanjutan untuk semua.

Untuk mendukung tercapainya target tersebut, perlu dilakukan prediksi dan klasifikasi kualitas air. Dengan memanfaatkan metode *machine learning* seperti *K-Nearest Neighbor* (K-NN) dan *Multinomial Logistic Regression*, data kualitas air dapat diolah untuk menghasilkan prediksi yang akurat. Pendekatan ini memungkinkan pemantauan kualitas air secara efektif dan dapat membantu dalam mengambil langkah preventif untuk mengatasi penurunan kualitas air sebelum berdampak lebih jauh pada kesehatan masyarakat, khususnya di Jawa Timur.

MASALAH

Masalah utama yang dihadapi berdasarkan pemaparan di atas adalah buruknya kualitas air di Jawa Timur. Kondisi ini memerlukan pemantauan kualitas air yang lebih efektif untuk mengidentifikasi dan mengatasi penurunan kualitas air. Dengan pemantauan yang tepat, langkah-langkah perbaikan dapat diambil untuk meningkatkan kualitas air,

sehingga masyarakat Jawa Timur dapat memiliki air bersih yang mendukung kesehatan mereka. Implementasi teknologi seperti *machine learning* dapat membantu dalam memprediksi dan mengklasifikasikan kualitas air, memastikan tindakan preventif dapat dilakukan secara tepat waktu.

KAJIAN PUSTAKA

Studi Terdahulu

Penelitian pertama yang dilakukan oleh (Tangkelayuk & Maiola, 2022) membahas mengenai klasifikasi kualitas air menggunakan metode K-NN, *Naïve Bayes* dan *Decision Tree*. Studi ini bertujuan untuk melihat perbandingan proses klasifikasi untuk mengetahui metode mana yang paling akurat, dilihat dari tingkat akurasi yang paling tinggi. Hasilnya, metode K-NN memiliki tingkat akurasi paling tinggi, yaitu sebesar 86.88%, sedangkan metode *Decision Tree* memiliki tingkat akurasi sebesar 80.84% dan *Naïve Bayes* sebesar 63.60%. Sehingga dapat disimpulkan bahwa metode K-NN merupakan metode yang paling baik untuk klasifikasi data tersebut.

Penelitian kedua yang dilakukan oleh (Roshini et al., 2023) membahas mengenai prediksi kualitas air menggunakan *Water Quality Index* (WQI). Penelitian ini dilakukan dengan tiga jenis model, yaitu *Non-Deep Learning Based Linear Regression Model, Deep Learning Based Linear Regression Model,* dan *Logistic Regression Model*. Hasil penelitian menunjukkan bahwa *Logistic Regression Model* menghasilkan persentase akurasi terbesar, yaitu sebesar 99,14%.

Teknologi

Teknologi yang dipakai dalam penelitian ini adalah *Machine Learning,* Metode *K-Nearest Neighbor,* Metode *Multinomial Logistic Regression, Dataset, Confusion Matrix, Correlation Map,* dan CRISP-DM. *Machine Learning* (ML) atau pembelajaran mesin dapat didefinisikan sebagai penggunaan komputer dan algoritma matematika untuk belajar dari data dan membuat prediksi di masa depan (Goldberg & Holland dalam Roihan et al., 2020). *K-Nearest Neighbor* merupakan sebuah metode yang dapat digunakan untuk melakukan klasifikasi terhadap suatu objek berdasarkan data pembelajaran yang memiliki jarak paling dekat dengan objek tersebut (Yustanti, 2012). *Multinomial Logistic Regression* adalah metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (y) yang bersifat polikotomus multinomial atau dibagi menjadi lebih dari dua

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 16-24, 2024
ISSN: 2808-7771 online

kategori. *Dataset* adalah kumpulan data tersusun yang dapat dianalisa untuk menjawab pertanyaan, membuat prediksi, dan membuat model. Kualitas kumpulan data sangat penting untuk keberhasilan *machine learning* atau ilmu data apa pun (Fakhri & Winursito, 2024). *Confusion matrix* merupakan sebuah alat ukur matriks yang dipakai untuk mendapat keakuratan jumlah klasifikasi terhadap kelas untuk algoritma yang dipakai (Qardini et al., 2021). *Correlation map* atau peta korelasi adalah tabel yang menunjukkan koefisien korelasi antar variabel dalam sebuah *dataset*. CRISP-DM *(Cross Industry Standard Process for Data Mining)* adalah metode standar yang digunakan dalam pengelolaan *data mining* dan analisis data, terdiri dari enam fase utama yang saling berhubungan: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment* (Wirth & Hipp, 2000).

METODE PELAKSANAAN

Dalam penelitian ini, diterapkan metode *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dengan model *K-Nearest Neighbor* dan *Multinomial Logistic Regression*. Metode ini terdiri dari enam fase, yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment*.

Business Understanding

Melihat banyaknya sumber air di Indonesia dengan kualitas air yang masih kurang baik serta dampaknya yang mengancam kesehatan masyarakat, prediksi dan klasifikasi dirasa sangat penting untuk dilakukan. Oleh karena itu, dipilih sebuah *dataset* mengenai faktor-faktor yang mempengaruhi kualitas air di Jawa Timur. Tujuannya adalah untuk mengetahui variabel mana yang paling berpengaruh, sehingga dapat menjadi salah satu upaya meningkatkan kualitas air di Indonesia, khususnya Jawa Timur. Hal ini sekaligus mendukung tercapainya Tujuan Pembangunan Berkelanjutan nomor 6 yang menargetkan terjaminnya ketersediaan dan pengelolaan air bersih serta sanitasi yang berkelanjutan untuk semua. Selain itu, penelitian ini juga dilakukan untuk memprediksi rata-rata kualitas air di Jawa Timur menggunakan metode *K-Nearest Neighbor* dan *Multinomial Logistic Regression*.

Data Understanding

Dataset yang digunakan untuk penelitian ini adalah *dataset* 8 Fitur Kualitas Air Jawa Timur, Indonesia yang diperoleh melalui situs *Kaggle*. Menurut Widiharsa (2023), data ini

diambil pada tahun 2021 dari tiga titik observasi: Sungai Cangkir Tambangan, Sungai Muara Kali Tengah, dan Bendungan Sutami. *Dataset* ini terdiri dari sembilan kolom yang mencakup berbagai parameter kualitas air sebagai berikut di tabel 1:

Tabel 1. Variabel yang terdapat di *dataset*

Variabel	Keterangan	Tipe Data
Temperatur	Suhu air yang mengidentifikasi apakah air dalam keadaan panas atau dingin.	Interval
***Total Dissolved Solid* (TDS)**	Jumlah zat padat terlarut dalam air, termasuk ion organik, senyawa, dan ion koloid.	Ratio
***Total Suspended Solid* (TSS)**	Total padatan yang tersuspensi di dalam air yang terdiri dari bahan organik dan anorganik yang dapat disaring dengan kertas millipore berpori pori 0,45 μm.	Ratio
***Power of Hydrogen* (pH)**	Tingkat keasaman dan kebasaan dari sebuah larutan yang masih mengandung air.	Interval
***Biochemical Oxygen Demand* (BOD)**	Permintaan oksigen biokimia, yaitu jumlah oksigen yang dibutuhkan oleh mikroorganisme untuk menguraikan bahan organik dalam air.	Ratio
***Chemical Oxygen Demand* (COD)**	Permintaan oksigen kimia atau jumlah oksigen yang dibutuhkan untuk mengoksidasi bahan organik dan anorganik dalam air secara kimia.	Ratio
***Dissolved Oxygen* (DO)**	Jumlah kadar oksigen yang terlarut di dalam air.	Ratio
Curah Hujan	Jumlah air hujan yang jatuh dalam periode tertentu.	Ratio
Kelas	Parameter hasil klasifikasi dengan tingkat kualitas air yang dinilai dari satu hingga empat, berdasarkan standar kualitas air yang ditetapkan oleh pemerintah Indonesia.	Interval

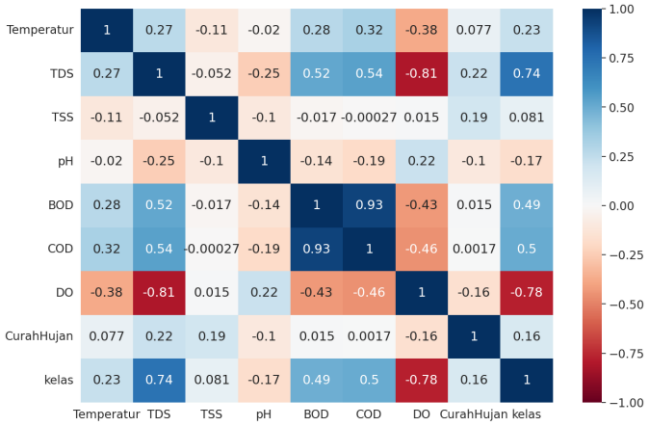
Berdasarkan PP nomor 22 tahun 2021 tentang Penyelenggaraan Perlindungan dan Pengelolaan Lingkungan Hidup, air yang diklasifikasikan sebagai berikut:

- Kelas Satu: Air yang dapat dipakai sebagai air baku air minum, dan peruntukan hal lain yang membutuhkan mutu air yang sama.
- Kelas Dua: Air yang digunakan sebagai prasarana rekreasi air, pembudidayaan ikan tawar, dan peruntukan hal lain yang membutuhkan mutu air yang sama.
- Kelas Tiga: Air yang dipakai untuk pembudidayaan ikan air tawar, peternakan, pengairan tanaman, dan peruntukan hal lain yang membutuhkan mutu air yang sama.

ISSN: 2808-7771 online

- Kelas Empat: Air yang digunakan untuk mengairi pertamanan, dan peruntukan hal lain yang membutuhkan mutu air yang sama.

Dataset ini menunjukkan informasi yang penting untuk memprediksi kondisi kualitas air di Jawa Timur. Untuk menggambarkan korelasi antar variabel, dapat divisualisasikan dengan sebuah *correlation map*. *Correlation Map* digunakan untuk mencari atribut-atribut dalam *dataset* yang dapat diambil untuk menjadi variabel independen yang paling berpengaruh terhadap variabel dependen. Gambar 2 merupakan *correlation map* dari *dataset* di atas:



Gambar 2. *Correlation Map*

Data Preparation

Data yang disediakan dalam *dataset* merupakan data yang sudah bersih dan tidak memuat *missing values*. Namun, terjadi inkonsistensi di mana pemilik *dataset* mendeskripsikan bahwa kualitas air diklasifikasikan ke dalam kelas satu hingga empat sesuai dengan peraturan pemerintah, sedangkan pada *dataset* terdapat data kelas dua hingga lima. Diasumsikan bahwa pengklasifikasian kualitas air yang benar adalah ke dalam kelas satu hingga empat, sehingga pada tahap ini dilakukan transformasi data untuk kelas 2 menjadi kelas 1, kelas 3 menjadi kelas 2, kelas 4 menjadi kelas 3, dan kelas 5 menjadi kelas 4. Berdasarkan *correlation map* yang sudah dibuat, variabel yang diambil sebagai variabel independen hanya temperatur, TDS, BOD, COD, dan DO, karena memiliki nilai korelasi yang cukup kuat dengan variabel dependen kelas.

Modeling

Pada tahap ini, *machine learning* digunakan untuk menghitung perbandingan nilai data yang akan diproses. *Dataset* dibagi menjadi dua bagian, yaitu *data training* dan *data testing*, dengan perbandingan 80:20, di mana 80% dari *dataset* digunakan untuk *training* dan 20% digunakan untuk *testing*. Algoritma *K-Nearest Neighbor* (K-NN) dengan k = 5 dan *Multinomial Logistic Regression* diterapkan pada data yang telah dibagi ini untuk

menentukan algoritma mana yang menghasilkan akurasi tertinggi. Dengan membandingkan hasil akurasi dari kedua algoritma ini, kita dapat menentukan metode yang paling efektif untuk memprediksi kondisi air di Jawa Timur.

Evaluation

Tahap *evaluation* adalah tahap yang memastikan kualitas dan efektivitas dari model bisnis yang dibangun. Tahap ini membutuhkan performa model berdasarkan tujuan bisnis dan menentukan kesesuaian untuk diaplikasikan di dunia nyata. Dalam penelitian ini, *classification report* digunakan sebagai ukuran kuantitatif untuk performa model *data mining*. Contoh dari *classification report* adalah *confusion matrix* yang menunjukkan jumlah contoh yang diklasifikasikan dengan benar dan salah untuk setiap kelasnya. Selain itu, digunakan juga *evaluation metrics* yang memberikan gambaran mengenai *accuracy*, *recall*, *precision*, dan *F1-Score* untuk sebuah model.

Deployment

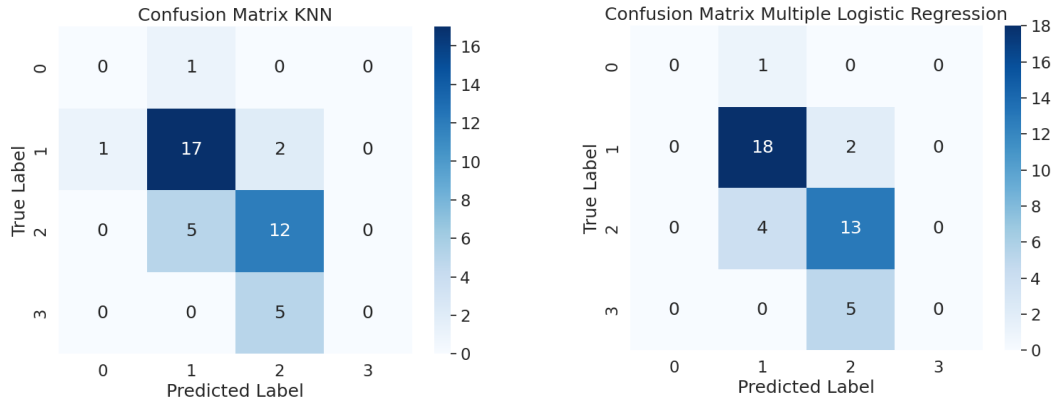
Tahapan ini dilakukan dengan membuat laporan jurnal menggunakan metode yang dihasilkan. Dengan menerapkan kedua algoritma ini, prediksi dapat dilakukan secara akurat. Sebagai contoh, menggunakan nilai rata-rata dari setiap variabel independen: Prediksikan kelas air dengan suhu 28.8°C, TDS 404.4 mg/L, BOD 21.43 mg/L, COD 92.53 mg/L, dan DO 3.7 mg/L. Setelah memasukkan semua variabel independen, hasil prediksi menunjukkan bahwa menurut algoritma *K-Nearest Neighbor* (K-NN), air termasuk dalam kelas 3, sedangkan menurut *Multinomial Logistic Regression*, air termasuk dalam kelas 2.

HASIL DAN PEMBAHASAN

Hasil pelaksanaan metode penelitian menunjukkan bahwa variabel yang paling mempengaruhi kelas kualitas air di Jawa Timur adalah *Dissolved Oxygen* (DO) dan *Total Dissolved Solid* (TDS). Kadar oksigen yang tinggi dalam air menunjukkan bahwa air tersebut layak digunakan dan mendukung kehidupan akuatik. Sebaliknya, kadar oksigen yang rendah menandakan air yang tercemar, yang dapat merusak ekosistem (Aruan & Siahaan, 2017). Sementara itu, tingginya kandungan TDS menunjukkan adanya polusi, seperti limbah kimia yang dapat membahayakan manusia (Makarim, 2022).

Untuk mengevaluasi kedua model, digunakan *confusion matrix* dan *evaluation metrics*. Gambar 3 adalah hasil untuk kedua model.

ISSN: 2808-7771 online



a b

Gambar 3. (a) *Confusion Matrix* model K-NN; (b) *Confusion Matrix* model *Multinomial Logistic Regression*

Dari hasil *confusion matrix* di atas, dapat diketahui bahwa model *Multinomial Logistic Regression* menghasilkan 2 *True Positive* lebih banyak dari model K-NN. Sementara itu, dari *evaluation metrics* terlihat bahwa nilai akurasi yang dihasilkan dari model *Multinomial Logistic Regression* juga lebih tinggi 0.05 jika dibandingkan dengan model K-NN. Oleh karena itu, dapat dikatakan bahwa metode *Multinomial Logistic Regression* lebih baik dalam memprediksi kualitas air di Jawa Timur. Selain itu, hasil tersebut menunjukkan bahwa model tidak wajib namun masih bisa ditingkatkan dengan pertimbangan untuk melanjutkan proyek atau membuat proyek baru.

KESIMPULAN

Penelitian ini mengidentifikasi bahwa variabel *Dissolved Oxygen* (DO) dan *Total Dissolved Solid* (TDS) memiliki pengaruh terbesar terhadap kualitas air di Jawa Timur, dengan menggunakan metode *machine learning K-Nearest Neighbor* (K-NN) dan *Multinomial Logistic Regression*. Hasil analisis menunjukkan bahwa tingginya kadar DO menandakan air yang layak dan mendukung kehidupan akuatik, sedangkan tingginya TDS menunjukkan polusi dan limbah berbahaya. Model *Multinomial Logistic Regression* memiliki akurasi lebih tinggi (0.72) dibandingkan K-NN (0.67), sehingga lebih efektif dalam memprediksi kualitas air. Untuk meningkatkan kualitas air, disarankan penggunaan teknologi pengolahan limbah seperti *reverse osmosis* atau elektrodialisis untuk mengurangi TDS, serta pembangunan air terjun buatan untuk meningkatkan DO. Kebijakan pengelolaan limbah yang ketat dan pengurangan pembuangan limbah oleh masyarakat juga

diperlukan. Implementasi sistem pemantauan kualitas air dengan *machine learning* dapat membantu tindakan preventif sebelum kondisi memburuk.

DAFTAR PUSTAKA

Badan Pusat Statistik Provinsi Jawa Timur. (n.d.). *Jumlah penduduk menurut jenis kelamin dan kabupaten/kota Provinsi Jawa Timur (jiwa), 2021-2023*. BPS Provinsi Jawa Timur. Retrieved June 6, 2024, from https://jatim.bps.go.id/indicator/12/375/1/jumlah-penduduk-provinsi-jawa-timur.html

Fakhri, A., & Winursito, Y. C. (2024, Februari). *Analisis penumpang kapal Titanic menggunakan Titanic dataset dengan bantuan pemrograman Python*. *Jurnal Sains Student Research, 2*(1), 539. https://doi.org/10.61722/jssr.v2i1.759

Makarim, D. F. R. (2022). *Wajib tahu, ini angka TDS yang layak untuk diminum*. Halodoc. Retrieved June 7, 2024, from https://www.halodoc.com/artikel/wajib-tahu-ini-angka-tds-yang-layak-untuk-diminum

Qardini, L., Seppewali, A., & Aina, A. (2021). *Decision tree dan AdaBoost pada klasifikasi penerima program bantuan sosial*. *Jurnal Inovasi Pendidikan (JIP), 2*(7), 1962. https://stp-mataram.e-journal.id/JIP/article/view/1046

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). *Pemanfaatan machine learning dalam berbagai bidang: Review paper*. *Indonesian Journal on Computer and Information Technology (IJCIT), 5*(1), 2. https://ejournal.bsi.ac.id/ejurnal/index.php/ijcit/article/view/7951

Roshini, B., Ranjitha, B., Sree, B. N., Lakshmi, B. B., & Sowmya, D. K. (2023, Maret). *Prediction of water quality using water quality index*. *International Research Journal of Modernization in Engineering Technology and Science, 05*(03), 756-760. https://www.doi.org/10.56726/IRJMETS34194

Tangkelayuk, A., & Mailoa, E. (2022, Juni). *Klasifikasi kualitas air menggunakan metode KNN, Naïve Bayes dan decision tree*. *Jurnal Teknik Informatika dan Sistem Informasi, 9*(2), 1109-1118. https://jurnal.mdp.ac.id/index.php/jatisi/article/download/2048/785/

Yustanti, W. (2012, Juli). *Algoritma K-Nearest Neighbour untuk memprediksi harga jual tanah*. *Jurnal Matematika, Statistika, & Komputasi, 9*(1), 60-61. https://journal.unhas.ac.id/index.php/jmsk/article/download/3399/1936

Widiharsa, P. (2023). *Dataset 8 fitur kualitas air Jawa Timur, Indonesia (Version 1)* [Dataset]. https://www.kaggle.com/datasets/prasetyawidiharsa/dataset-8-fitur-kualitas-air-jawa-timur-indonesia

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-39).