

**KLASIFIKASI *DATA MINING* UNTUK MEMPREDIKSI KUALITAS AIR
YANG DAPAT DIKONSUMSI MENGGUNAKAN ALGORITMA
*RANDOM FOREST***

SKRIPSI

Diajukan Untuk Penelitian Skripsi Guna Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer (S.Kom.)
Program Studi Sistem Informasi UN PGRI Kediri



OLEH :

SAYLENDRA ARGA WARDANI

NPM : 2113030052

FAKULTAS TEKNIK DAN ILMU KOMPUTER (FTIK)
UNIVERSITAS NUSANTARA PERSATUAN GURU REPUBLIK INDONESIA
UN PGRI KEDIRI

2025

Skripsi oleh :

SAYLENDRA ARGA WARDANI

NPM : 2113030052

Judul:

**KLASIFIKASI *DATA MINING* UNTUK MEMPREDIKSI KUALITAS AIR
YANG DAPAT DIKONSUMSI MENGGUNAKAN ALGORITMA
*RANDOM FOREST***

Telah disetujui untuk diajukan Kepada
Panitia Ujian/Sidang Skripsi Program Studi Sistem Informasi
FTIK UN PGRI Kediri

Tanggal: 2 Juli 2025

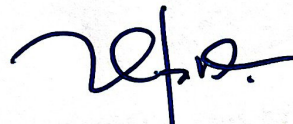
Pembimbing I



Sucipto, M. Kom

NIDN. 0721029101

Pembimbing II



Anita Sari Wardani. M. Kom

NIDN. 0713018402

Skripsi oleh :
SAYLENDRA ARGA WARDANI
NPM : 2113030052

Judul
**KLASIFIKASI *DATA MINING* UNTUK MEMPREDIKSI KUALITAS AIR
YANG DAPAT DIKONSUMSI MENGGUNAKAN ALGORITMA
*RANDOM FOREST***

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi
Program Studi Sistem Informasi FTIK UN PGRI Kediri
Tanggal: 10 Juli 2025

Dan Dinyatakan telah Memenuhi Persyaratan

Panitia Penguji :

1. Ketua Penguji : Sucipto, M. Kom

[.....]

2. Penguji 1 : Dwi Harini, M.M

[.....]

3. Penguji 2 : Anita Sari Wardani, M.Kom

[.....]



Mengetahui,

Dekan Fakultas Teknik dan Ilmu Komputer

Dr. Sulistiono, M.Si
NIDN. 0007076801

PERNYATAAN

Yang bertanda tangan di bawah ini saya,

Nama	: Saylendra Arga Wardani
Jenis Kelamin	: Laki-laki
Tempat/tgl. lahir	: Kediri / 16 April 2003
NPM	: 2113030052
Fak/Prodi.	: FTIK/ S1-Sistem Informasi

menyatakan dengan sebenarnya, bahwa dalam Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya tidak terdapat karya tulis atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara sengaja dan tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Kediri, 2 Juli 2025

Yang Menyatakan

SAYLENDRA ARGAS WARDANI

NPM: 2113030052

MOTTO

“Jika Kamu Tidak Menemukan Jalan, Maka Buatlah Jalanmu Sendiri”

~ Roronoa Zoro ~

"Jika Kamu Tidak Mengambil Resiko, Maka Kamu Tidak Akan Dapat
Menciptakan Masa Depan Mu Sendiri"

~ Monkey D Luffy ~

ABSTRAK

Saylendra Arga Wardani: Klasifikasi *Data Mining* Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi Menggunakan Algoritma *Random Forest*, Skripsi, Sistem Informasi, Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri, 2025.

Kata Kunci: *Data Mining*, *Random Forest*, Kualitas Air, Klasifikasi, *K-Fold Cross Validation*, *Feature Importance*.

Ketersediaan air minum yang layak merupakan salah satu faktor krusial dalam menjamin kesehatan masyarakat. Penilaian terhadap potabilitas air memerlukan pendekatan analitik yang andal, terutama dalam memanfaatkan data parameter fisik dan kimia air. Beberapa studi terdahulu belum mengoptimalkan proses validasi model, sehingga berisiko menghasilkan estimasi kinerja yang kurang representatif.

Penelitian ini bertujuan untuk mengkaji performa algoritma *Random Forest* dalam mengklasifikasikan kelayakan air minum serta mengidentifikasi fitur-fitur yang paling berpengaruh melalui analisis *feature importance*. Dataset yang digunakan adalah “*Water Quality and Potability*” yang tersedia secara publik melalui platform *Kaggle*. Tahapan praproses data meliputi imputasi terhadap nilai yang hilang, deteksi dan penghapusan *outlier* menggunakan metode *Interquartile Range (IQR)*, normalisasi data dengan *StandardScaler*, serta penyeimbangan distribusi kelas menggunakan metode *upsampling*. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score*, serta divalidasi menggunakan teknik *Stratified K-Fold Cross Validation* dengan 10 lipatan.

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* mampu menghasilkan akurasi sebesar 81,68% pada data uji dan rata-rata 78,23% pada evaluasi menggunakan validasi silang. Kinerja ini lebih unggul dibandingkan studi sebelumnya yang belum menerapkan validasi secara menyeluruh. Selain itu, fitur *Hardness*, *pH*, *Chloramines*, dan *Solids* teridentifikasi sebagai atribut yang paling signifikan dalam memengaruhi klasifikasi potabilitas.

Penerapan algoritma *Random Forest* yang dikombinasikan dengan validasi *K-Fold* terbukti efektif dalam membangun model klasifikasi kualitas air yang akurat dan andal. Keterbatasan penelitian terletak pada keterbatasan jumlah dan variasi data. Penelitian selanjutnya direkomendasikan untuk menguji pendekatan serupa pada dataset yang lebih besar atau dengan algoritma pembelajaran mesin lainnya.

KATA PENGANTAR

Puji Syukur Kami panjatkan Kehadirat Allah Tuhan Yang Maha Kuasa, karena hanya atas perkenan-Nya penyusunan skripsi ini dapat diselesaikan. Skripsi dengan judul “Klasifikasi *Data Mining* Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi Menggunakan Algoritma *Random Forest*” ini ditulis guna memenuhi sebagian syarat untuk memperoleh gelar Sarjana Komputer, pada Program Studi Sistem Informasi Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri. Pada kesempatan ini diucapkan terima kasih dan penghargaan yang setulus-tulusnya kepada:

1. Dr. Zainal Afandi, M.Pd., selaku Rektor Universitas Nusantara PGRI Kediri
2. Dr. Sulistiono, M.Si. selaku Dekan Fakultas Teknik dan Ilmu Komputer Universitas Nusantara PGRI Kediri.
3. Sucipto, M. Kom., selaku Ketua Prodi Sistem Informasi Fakultas Teknik dan Ilmu Komputer Universitas Nusantara PGRI Kediri
4. Sucipto, M. Kom., selaku Dosen Pembimbing I
5. Anita Sari Wardani M.Kom., selaku Dosen Pembimbing II
6. Orang tua tercinta atas motivasi, dukungan, serta doa yang tak henti-hentinya
7. Teman-teman seperjuangan yang selalu memberikan dukungan dan semangat
8. Ucapan terima kasih juga disampaikan kepada pihak-pihak lain yang tidak dapat disebutkan satu persatu, yang telah banyak membantu menyelesaikan Skripsi ini

Disadari bahwa skripsi ini masih banyak kekurangan, maka diharapkan tegur sapa, kritik dan saran-saran, dari berbagai pihak sangat diharapkan.

Kediri, 2 Juli 2025

SAYLENDRA ARGA WARDANI

NPM: 2113030052

DAFTAR ISI

Halaman Sampul	i
Halaman Persetujuan.....	ii
Halaman Pengesahan	iii
Halaman Pernyataan.....	iv
Halaman Motto.....	v
ABSTRAK	vi
KATA PENGANTAR.....	vii
DAFTAR ISI	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xi
DAFTAR LAMPIRAN	xii
BAB I PENDAHULUAN	1
A. Latar Belakang Masalah.....	1
B. Batasan Masalah.....	3
C. Rumusan Masalah	3
D. Tujuan Penelitian.....	3
E. Manfaat Penelitian	4
BAB II KAJIAN PUSTAKA	5
A. Kajian Teori.....	5
B. Kajian Penelitian Terdahulu	15
C. Kerangka Berpikir.....	18
BAB III METODE PENELITIAN	21
A. Desain Penelitian.....	21
B. Alat dan Bahan.....	24
C. Populasi dan Sampel/ Objek Penelitian	25
D. Prosedur Penelitian.....	26
E. Analisis Hasil	28
BAB IV HASIL DAN PEMBAHASAN.....	30
A. <i>Bussiness Understanding</i>	30

B. <i>Data Understanding</i>	31
C. <i>Data Preparation</i>	48
D. <i>Modeling</i>	56
E. Evaluasi.....	65
F. <i>Deployment</i>	72
G. Pembahasan Hasil Temuan.....	74
BAB V PENUTUP	78
A. Kesimpulan	78
B. Saran.....	78
DAFTAR PUSTAKA	80
LAMPIRAN	84

DAFTAR GAMBAR

Gambar 2.1. Visualisasi <i>Random Forest</i>	9
Gambar 2.3. Ilustrasi Pembagian Data Uji.....	11
Gambar 2.8. Kerangka Berfikir.....	18
Gambar 3. 1 Alur <i>CRISP-DM</i>	21
Gambar 3.2. Alur Penelitian.....	26
Gambar 4.1. Distribusi Data Atribut <i>pH</i>	33
Gambar 4.2. Distribusi Data Atribut <i>Hardness</i>	35
Gambar 4.3. Distribusi Data Atribut <i>Solids</i>	36
Gambar 4.4. Distribusi Data Atribut <i>Chloramines</i>	38
Gambar 4.5. Distribusi Data Atribut <i>Sulfate</i>	39
Gambar 4.6. Distribusi Data Atribut <i>Conductivity</i>	41
Gambar 4.7. Distribusi Data Atribut <i>Organic Carbon</i>	42
Gambar 4.8. Distribusi Data Atribut <i>Trihalomethanes</i>	44
Gambar 4.9. Distribusi Data Atribut <i>Turbidity</i>	45
Gambar 4.10. Distribusi Data Atribut <i>Potability</i>	47
Gambar 4.11. Visualisasi Data Sebelum Hapus <i>Outlier</i>	52
Gambar 4.12. Visualisasi Data Sesudah Hapus <i>Outlier</i>	53
Gambar 4.13. Jumlah Data Sesudah Hapus <i>Outlier</i>	53
Gambar 4.14. Jumlah Data Potability Setelah <i>UpSampling</i>	54
Gambar 4.15. Simulasi Pohon Keputusan <i>Random Forest</i>	59
Gambar 4.16. Hasil <i>Feature Importance</i>	62
Gambar 4.17. <i>Confusion Matrix</i>	65

DAFTAR TABEL

Tabel 2.1. <i>Confusion Matrix</i>	12
Tabel 4.1. Tampilan Dataset	31
Tabel 4.2. Keterangan Isi Dataset	32
Tabel 4.3. Data Sebelum Penanganan Nilai Hilang	50
Tabel 4.4. Hasil Penanganan Nilai Hilang	51
Tabel 4.5. Data Sebelum Normalisasi	55
Tabel 4.6. Data Setelah Normalisasi	55
Tabel 4.7. Data Latih	57
Tabel 4.8. Data Uji	58
Tabel 4.9. Hasil Nilai <i>Feature Importance</i>	63
Tabel 4.10. Evaluasi Hasil.....	66
Tabel 4.11. <i>Classification Report</i>	67
Tabel 4.12. Hasil <i>K-Fold Cross Validation</i>	71
Tabel 4.13. Perbandingan Metrik Evaluasi	76

DAFTAR LAMPIRAN

Lampiran 1 Kartu Bimbingan Skripsi	86
Lampiran 2 Surat Keterangan Bebas <i>Similarity</i>	87
Lampiran 3 Bukti Halaman Awal cek <i>Similarity</i>	88
Lampiran 4 Bukti Submit Artikel/Loa/Terbit	89

BAB I PENDAHULUAN

A. Latar Belakang Masalah

Air adalah sumber daya alam yang sangat penting untuk semua makhluk hidup, terutama bagi manusia. Semua bagian masyarakat berusaha untuk mendapatkan sumber air yang terbaik untuk memenuhi kebutuhan mereka sehari-hari. Air merupakan sumber daya yang dimanfaatkan untuk berbagai kebutuhan manusia seperti untuk di konsumsi, perumahan, irigasi, peternakan, perikanan, pembangkit listrik, transportasi, industri, dan rekreasi. Setiap daerah membutuhkan air yang aman untuk dikonsumsi, tetapi kualitas air semakin menurun di beberapa tempat, terutama untuk memenuhi kebutuhan manusia akan air minum (Fitriyono et al., 2024).

Seiring berkembangnya teknologi informasi, penerapan metode komputasi cerdas seperti *data mining* menjadi salah satu solusi untuk mengatasi permasalahan dalam mengolah data kualitas air secara cepat dan akurat. *Data mining* adalah istilah yang mengacu pada penggabungan metode yang bertujuan untuk menemukan pola yang tidak diketahui dalam data yang telah dikumpulkan (Tarigan et al., 2022). *Data mining* adalah proses semi-otomatis yang menggunakan matematika, statistik, kecerdasan buatan, dan pembelajaran mesin untuk mengekstraksi dan mengidentifikasi informasi (Zai, 2022) serta mampu mengekstrak pola-pola tersembunyi dari data historis, yang kemudian dapat digunakan sebagai landasan untuk melakukan prediksi dan mengambil tindakan yang lebih informatif (Nugroho & Husin, 2022). Melalui data mining, informasi penting dari data historis dapat diekstraksi guna membantu proses pengambilan keputusan yang lebih baik dan berbasis data.

Salah satu algoritma yang sering digunakan dalam *data mining* untuk prediksi adalah *Random Forest*. *Random forest* adalah metode dalam pembelajaran *ensemble* yang menciptakan model prediksi dengan menggabungkan banyak model pohon keputusan yang dibuat secara acak (Nugroho, 2022). Dalam membangun model, dataset pelatihan dibagi menjadi beberapa subset untuk membentuk masing-masing pohon keputusan. Setelah model terbentuk, dilakukan pengujian data untuk menilai kinerja setiap pohon keputusan. Setiap pohon kemudian memberikan

suaranya berdasarkan model yang telah dibuat, dan prediksi akhir ditentukan oleh suara terbanyak (Nugroho et al., 2021). Keunggulan *Random Forest* dalam menangani prediksi yaitu dapat menghasilkan performa akurasi tinggi dan kemampuan generalisasi model pada data baru dengan meminimalkan risiko *overfitting*. Kelebihan lain dari algoritma ini adalah kemampuannya dalam mengukur pentingnya setiap fitur (*feature importance*), yang berguna untuk mengetahui variabel mana yang paling berpengaruh terhadap hasil klasifikasi. Oleh karena itu, algoritma *Random Forest* berpotensi besar untuk digunakan dalam klasifikasi potabilitas air.

Beberapa penelitian sebelumnya telah menggunakan algoritma *Random Forest* dalam mengklasifikasikan kualitas air minum. Salah satunya adalah Penelitian terdahulu yang dilakukan oleh *Christian Yefta, dkk (2022)* dalam penelitiannya yang berjudul “*Prediksi Kualitas Air Menggunakan Algoritma Naïve Bayes Dan Random Forest*”, dari hasil penelitian tersebut diperoleh pengukuran model menggunakan tabel *confusion matrix*, didapatkan hasil akurasi sebesar 79% untuk algoritma *random forest* dan algoritma *naïve bayes* menghasilkan akurasi sebesar 55% (Christian et al., 2022). Namun demikian, penelitian tersebut belum mengevaluasi model secara menyeluruh dengan teknik validasi silang *seperti K-Fold Cross Validation*. Hal ini menyebabkan tingkat akurasi dan kemampuan generalisasi model terhadap data baru belum dapat dipastikan secara optimal.

Selain itu, belum banyak penelitian yang mengidentifikasi dan menganalisis fitur-fitur penting (*feature importance*) yang memengaruhi tingkat potabilitas air. Padahal, informasi ini dapat memberikan wawasan penting bagi pengambil kebijakan dalam menetapkan standar kualitas air dan strategi penanganannya. Kurangnya pendekatan evaluatif dan analitis dalam penelitian sebelumnya menciptakan celah yang dapat diisi oleh penelitian ini.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk mengembangkan model klasifikasi potabilitas air menggunakan algoritma *Random Forest* yang dievaluasi secara menyeluruh dengan metode evaluasi *K-Fold Cross Validation*. Penelitian ini juga bertujuan untuk mengidentifikasi variabel-variabel yang paling signifikan terhadap potabilitas air melalui analisis *feature importance*.

Diharapkan hasil dari penelitian ini dapat menjadi dasar dalam pengambilan keputusan yang lebih akurat dan mendukung upaya pemenuhan akses air minum yang aman bagi masyarakat.

B. Batasan Masalah

Penelitian ini memiliki batasan-batasan masalah yang meliputi sebagai berikut:

1. Algoritma yang digunakan dalam penelitian ini dibatasi pada algoritma *Random Forest*.
2. Dataset yang digunakan dalam penelitian ini adalah *Water Quality and Potability* yang diunduh dari platform *Kaggle* (<https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability>).
3. Prediksi kualitas air dalam penelitian ini hanya mencakup variabel-variabel yang tersedia pada dataset yang digunakan.

C. Rumusan Masalah

Berdasarkan Batasan masalah yang telah diuraikan, rumusan permasalahan pada penelitian ini, disimpulkan yaitu:

1. Berapa akurasi yang diperoleh dari penerapan evaluasi model menggunakan *K-Fold Cross Validation* pada algoritma *Random Forest* dalam memprediksi kualitas air yang layak dikonsumsi?
2. Atribut-atribut mana saja yang paling signifikan mempengaruhi kualitas air berdasarkan hasil analisis menggunakan algoritma *Random Forest*, dan bagaimana pengaruhnya terhadap hasil pengujian ulang model pada *data training* dan *data testing*?

D. Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Mengevaluasi kinerja algoritma *Random Forest* dalam memprediksi kualitas air yang layak dikonsumsi dengan menggunakan Teknik evaluasi model *K-Fold Cross Validation* untuk menentukan tingkat akurasi model.
2. Mengidentifikasi atribut-atribut yang paling signifikan mempengaruhi kualitas air berdasarkan hasil analisis menggunakan algoritma *Random Forest*, serta mengevaluasi pengaruhnya terhadap hasil pengujian ulang model pada *data training* dan *data testing*.

E. Manfaat Penelitian

Manfaat yang didapat dari penelitian ini yaitu sebagai berikut:

1. Memberikan wawasan yang lebih mendalam tentang efektivitas algoritma *Random Forest* melalui evaluasi menggunakan *K-Fold Cross Validation*, serta mengidentifikasi atribut-atribut signifikan yang mempengaruhi kualitas air untuk meningkatkan akurasi dan kemampuan generalisasi model dalam memprediksi potabilitas air.
2. Berkontribusi pada pengembangan ilmu pengetahuan di bidang *data mining* dan pembelajaran mesin terkait kualitas air, sekaligus menjadi dasar untuk penelitian lanjutan dan penerapan model prediksi yang lebih kompleks atau algoritma lainnya.

DAFTAR PUSTAKA

- A'yunnisa, N., Salim, Y., & Azis, H. (2022). Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab. *Indonesian Journal of Data and Science (IJODAS)*, 3(3), 115–121. <https://doi.org/https://doi.org/10.56705/ijodas.v3i3.54>
- Abadi, M. F. (2023). KLASIFIKASI LAMA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR. In *Teknologipintar.org* (Vol. 2, Issue 12). Teknologipintar.org
- Azmi, B. N., Hermawan, A., & Avianto, D. (2022). Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma K-Nearest Neighbor dan Logistic Regression. *Jurnal Sistem Dan Teknologi Informasi*, 7(2), 94–103. <https://doi.org/https://doi.org/10.32528/justindo.v7i2.8190>
- Cahyo, D. A. (2023). METODE NAIVE BAYES UNTUK KLASIFIKASI MASA STUDI SARJANA. In *Teknologipintar.org* (Vol. 3, Issue 4).
- Christian, Y., Putra, J., Winata, A., Ricky, N., Jeonanto, R., & Artikel, H. (2022). *PREDIKSI KUALITAS AIR MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN RANDOM FOREST*. <https://doi.org/10.24269/jkt.v6i2.1313>
- Darmawan, I., Muhammad Fatchan, & Andri Firmansyah. (2024). Classification of Drinking Water Potability With Artificial Neural Network Algorithm. *International Journal of Integrated Science and Technology*, 2(5), 506–515. <https://doi.org/10.59890/ijist.v2i5.1874>
- Farrel Budiantara, A., & Budihartanti, C. (2020). IMPLEMENTASI DATA MINING DALAM MANAJEMEN INVENTORY PADA PT. MASTERSYSTEM INFOTAMA MENGGUNAKAN METODE ALGORITMA APRIORI. *Jurnal PROSISKO*, 7(1), 26–31. <https://doi.org/https://doi.org/10.30656/prosisko.v7i1.2130>
- Faruqziddan, M., Herdika Septa Aulia, E., Dini Azzahra, S., **Ristyawan, A., & Daniati, E. (2024)**. Klasifikasi Risiko Kambuhnya Kanker Tiroid Menggunakan Algoritma Random Forest. In *Agustus* (Vol. 8). Online. <https://doi.org/https://doi.org/10.29407/inotek.v8i1.4912>

- Fernando, D., & Guntara, R. G. (2024). Model Klasifikasi Penyebab Turnover Karyawan Menggunakan Kerangka Kerja CRISP-DM. *J-INTECH (Journal of Information and Technology)*. <https://doi.org/https://doi.org/10.32664/j-intech.v12i02.1502>
- Fitri Boy, A., Yakub, S., Azmi, Z., & Triguna Dharma, S. (2022). IMPLEMENTASI DATA MINING PADA PENGATURAN DISTRIBUSI BARANG DENGAN MENGGUNAKAN ALGORITMA FP-GROWTH. In *Journal of Science and Social Research* (Issue 2). <https://doi.org/https://doi.org/10.54314/jssr.v5i2.947>
- Fitriono, D., Wardani, S. A., Nizar, M., Al Varuq, B., **Ristyan, A., & Daniati, E. (2024).** Perbandingan Metode Algoritma Decision Tree dan K-Nearest Neighbors untuk Memprediksi Kualitas Air yang dapat dikonsumsi. In *Agustus* (Vol. 8). Online. <https://doi.org/https://doi.org/10.29407/inotek.v8i1.4978>
- Jimmy, Dwi Yulianto, L., Heni Hermaliani, E., & Kurniawati, L. (2023). Penerapan Machine Learning Dalam Analisis Stadium Penyakit Hati Untuk Proses Diagnosis dan Perawatan. *RESOLUSI: Rekayasa Teknik Informatika Dan Informasi*, 3(4), 170–180. <https://doi.org/https://doi.org/10.30865/resolusi.v3i4.709>
- Khairudin. (2024). PENGENALAN BAHASA PEMROGRAMAN PYTHON UNTUK MENINGKATKAN KOMPETENSI YATIM PIATU RW 01 KELURAHAN PANUNGGAN KEC PINANG. *Praxis : Jurnal Pengabdian Kepada Masyarakat*, 4, 36–43. <https://pijarpemikiran.com/index.php/praxis/article/view/716>
- Krisna Amarya, T., Candra Andy, A. G., Achmad, R., **Daniati, E., & Ristyan, A. (2024).** Analisa Perbandingan Algoritma Classification Berdasarkan Komposisi Label. In *Agustus* (Vol. 8). Online. <https://doi.org/https://doi.org/10.29407/inotek.v8i1.4906>
- Nanda Aulia Sofiah, Fanny Olivia, & Jambak, M. I. (2024). Predictive Analytics for Water Safety: Data Mining and Supervised Learning in Potability Classification. *The Indonesian Journal of Computer Science*, 13(4), 5066–5077. <https://doi.org/10.33022/ijcs.v13i4.3884>
- Nisa Sofia Amriza, R., Supriyadi, D., Jl Panjaitan No, P. DI, Purwokerto Selatan, K., Banyumas, K., & Tengah, J. (2021). Komparasi Metode Machine Learning dan

- Deep Learning untuk Deteksi Emosi pada Text di Sosial Media. In *Jurnal JUPITER* (Vol. 13, Issue 2). <https://doi.org/https://doi.org/10.5281/3603.jupiter.2021.10>
- Nugroho, A.** (2022). Analisa Splitting Criteria Pada Decision Tree dan Random Forest untuk Klasifikasi Evaluasi Kendaraan. *JSITIK: Jurnal Sistem Informasi Dan Teknologi Informasi Komputer*, 1(1), 41–49. <https://doi.org/10.53624/jsitik.v1i1.154>
- Nugroho, A., Fanani, A. Z., & Shidik, G. F.** (2021). *Evaluation of Feature Selection Using Wrapper For Numeric Dataset With Random Forest Algorithm*. IEEE. <https://doi.org/10.1109/iSemantic52711.2021.9573249>
- Nugroho, A., & Husin, A. (2022). Analisis Performa Random Forest Menggunakan Normalisasi Atribut. In *Januari* (Vol. 11, Issue 1). <https://doi.org/https://doi.org/10.32520/stmsi.v11i1.1681>
- Purnamawati, A., Nugroho, W., Putri, D., & Hidayat, W. F. (2020). Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN. *InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan*, 5(1), 212–215. <https://doi.org/10.30743/infotekjar.v5i1.2934>
- Ruswanti, D., Susilo, D., & Riani, R. (2024). Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45. *Go Infotech: Jurnal Ilmiah STMIK AUB*, 30(1), 111–121. <https://doi.org/10.36309/goi.v30i1.266>
- Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, 14(1), 155–162. <https://doi.org/10.35970/infotekmesin.v14i1.1751>
- Savitri, L., & Nursalim, R. (2023). *Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning*. <https://doi.org/https://doi.org/10.33369/diophantine.v2i01.28260>
- Tarigan, P. M. S., Hardinata, J. T., Qurniawan, H., Safii, M., & Winanjaya, R. (2022). Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan

- Persediaan Barang. *Jurnal Janitra Informatika Dan Sistem Informasi*, 2(1), 9–19.
<https://doi.org/10.25008/janitra.v2i1.142>
- Triono, A., Setia Budi, A., & Abdillah, R. (2023). IMPLEMENTASI PERETASAN SANDI VIGENERE CHIPHER MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON. In *Jurnal JOCOTIS-Journal Science Informatica and Robotics E-ISSN : xxxx-xxxx* (Vol. 1, Issue 1).
- Utomo, D. P., & Mesran. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 4(2), 437–444.
<https://doi.org/10.30865/mib.v4i2.2080>
- Wardhani, D., Astuti, R., & Saputra, D. D. (2024). Optimasi Feature Selection Text Mining: Stemming dan Stopword. *INNOVATIVE: Journal Of Social Science Research*, 4, 7537–7548.
<https://doi.org/https://doi.org/10.31004/innovative.v4i1.8759>
- Weningtyas, A., & Widuri, E. (2022). Pengelolaan Sumber Daya Air Berbasis Kearifan Lokal Sebagai Modal Untuk Pembangunan Berkelanjutan. *Volksgeist: Jurnal Ilmu Hukum Dan Konstitusi*, 5(01), 129–144.
<https://doi.org/doi.org/10.24090/volksgeist.v5i1.6074>
- Wibowo, M., & Djafar, Muh. R. F. (2023). Perbandingan Metode Klasifikasi Untuk Deteksi Stress Pada Mahasiswa di Perguruan Tinggi. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7(1), 153–159.
<https://doi.org/10.30865/mib.v7i1.5182>
- Wulandari, V., Sari, W. J., Alfian, Z., Legito, L., & Arifianto, T. (2024). Implementasi Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(2), 710–718. <https://doi.org/10.57152/malcom.v4i2.1229>
- Zai, C. (2022). IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA. In *Portaldata.org* (Vol. 2, Issue 3). Portaldata.or