Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

ISSN 2355-5920 | e-ISSN 2655-1845
https://ejournal.unib.ac.id/pseudocode/index

Klasifikasi *Burnout* Mahasiswa menggunakan Algoritma *Random Forest* dengan Analisis *Feature Importance*

Andrean Yuriko Wibisono, Abd. Hadi*

Fakultas Teknologi dan Desain, Institut Asia Malang, Jl. Rembuksari No.1 A, Mojolangu, Kec. Lowokwaru, Malang 65113, Indonesia

Informasi Naskah:
Diterima: 18 Januari 2026/ Direview: 24 Januari 2026/ Direvisi: 06 Februari 2026/ Disetujui Terbit: 06 Februari 2026
DOI: 10.33369/pseudocode.13.1.54-62
*Korespondensi: hadi@asia.ac.id

Abstract

Prolonged stress in higher education environments can lead to emotional, physical, and mental exhaustion among students, negatively affecting concentration, academic performance, and mental health. This study aims to classify student burnout levels using the Random Forest algorithm and to analyze model performance as well as dominant contributing factors through a feature importance approach. The dataset was obtained from the Kaggle platform and consisted of 1,100 samples covering psychological, physiological, environmental, academic, and social variables. The research methodology included data preprocessing, data balancing using the Synthetic Minority Over-sampling Technique (SMOTE), and data splitting with six scenarios: 70:30, 75:25, 78:22, 80:20, 83:17, and 87:13. Experimental results showed that all splitting scenarios produced relatively stable performance, with the 80:20 split achieving the highest accuracy of 89%, while other scenarios ranged from 85% to 87%. Feature importance analysis indicated that physiological factors, particularly blood_pressure, had the most significant contribution to student burnout classification.

Keywords: Data Mining; Feature Importance; Machine Learning; Random Forest; Student Burnout

1. Pendahuluan

Tekanan akademik yang berlangsung secara terus-menerus, ditambah dengan tuntutan kehidupan sosial dan lingkungan di perguruan tinggi, dapat memicu kondisi kelelahan emosional, fisik, dan mental pada mahasiswa. Kondisi ini dikenal sebagai *academic burnout* dan sering dikaitkan dengan penurunan konsentrasi, menurunnya performa akademik, serta gangguan kesehatan mental. Penelitian yang dilakukan oleh Wardhana et al. [1] menunjukkan bahwa burnout pada mahasiswa dipengaruhi oleh kombinasi faktor internal dan eksternal, termasuk tuntutan akademik, kondisi psikologis, serta lingkungan sosial. Apabila tidak terdeteksi sejak dini, burnout berpotensi berdampak pada meningkatnya stres akademik dan risiko putus studi.

Fenomena *burnout* pada mahasiswa juga menjadi perhatian serius dalam konteks pendidikan tinggi di Indonesia. Penelitian yang dilakukan oleh Damayanti dan Aulia [2] menunjukkan bahwa mahasiswa Indonesia mengalami *burnout* akademik yang ditandai oleh kelelahan emosional, menurunnya minat belajar, serta perasaan tertekan terhadap tuntutan akademik. Kondisi ini tidak hanya dipengaruhi oleh beban perkuliahan, tetapi juga oleh faktor psikologis dan sosial yang melekat pada kehidupan mahasiswa. Temuan tersebut sejalan dengan kajian yang dilakukan oleh Wardhana et al. [1], yang menyimpulkan bahwa *burnout* mahasiswa di Indonesia bersifat multidimensional dan dipengaruhi oleh interaksi antara faktor internal dan eksternal. Sebagai upaya mitigasi, pemanfaatan teknologi cerdas seperti media konsultasi kesehatan mental berbasis *Natural Language Processing* (NLP) kini mulai dikembangkan untuk membantu mendeteksi gejala depresi dan stres mahasiswa secara dini [3]. Oleh karena itu, pendekatan analitis yang mampu menangkap kompleksitas faktor-faktor tersebut diperlukan agar upaya pencegahan dan penanganan *burnout* dapat dilakukan secara lebih komprehensif.

Seiring berkembangnya teknologi, pendekatan berbasis *machine learning* dan kecerdasan buatan mulai banyak dimanfaatkan untuk menganalisis permasalahan kesehatan mental, termasuk pengembangan sistem konsultasi otomatis untuk mendeteksi kecemasan dan stres [3]. Dalam konteks klasifikasi yang kompleks, Haddouchi dan Berrado [4] menjelaskan bahwa algoritma *Random Forest* memiliki keunggulan signifikan dalam menangani data berdimensi tinggi dan bersifat *nonlinier*, serta relatif tahan terhadap *overfitting* berkat mekanisme agregasi *bootstrap*. Lebih lanjut, meskipun *Random Forest* sering dikategorikan sebagai model *black box* karena kompleksitas struktur pohonnya, algoritma ini menyediakan mekanisme *feature importance* (VIPs). Mekanisme ini memungkinkan peneliti untuk mengidentifikasi kontribusi setiap variabel terhadap hasil prediksi, sehingga transparansi dan interpretabilitas model dapat ditingkatkan secara signifikan.

Namun, salah satu tantangan dalam penerapan klasifikasi pada data kesehatan mental adalah ketidakseimbangan kelas. Penelitian yang dilakukan oleh Erlin et al. [5] menunjukkan bahwa data tidak seimbang dapat menurunkan performa model klasifikasi jika tidak ditangani dengan tepat. Untuk

Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

mengatasi permasalahan tersebut, *Synthetic Minority Over-sampling Technique* (SMOTE) sering digunakan karena mampu meningkatkan representasi kelas minoritas secara sintetis tanpa menghilangkan karakteristik data asli. Oktaviani et al. [6] membuktikan bahwa kombinasi *Random Forest* dan SMOTE menghasilkan performa klasifikasi yang lebih stabil pada kasus stres dan tekanan akademik mahasiswa dibandingkan tanpa penyeimbangan data.

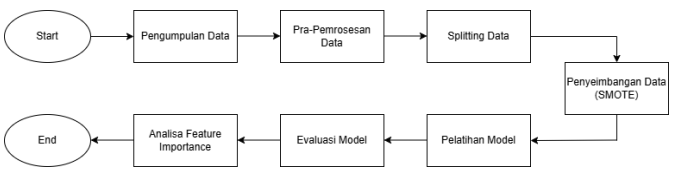
Meskipun berbagai penelitian telah menerapkan *Random Forest* dalam klasifikasi kondisi psikologis mahasiswa, sebagian besar studi masih berfokus pada evaluasi performa model, seperti akurasi dan *recall*, tanpa membahas secara mendalam faktor-faktor dominan yang memengaruhi *burnout*. Menurut Haddouchi dan Berrado [4], analisis *feature importance* seharusnya dimanfaatkan sebagai sarana interpretasi untuk menjembatani hasil komputasi dengan pengambilan keputusan di dunia nyata. Selain itu, kajian yang mengintegrasikan faktor fisiologis secara eksplisit dalam klasifikasi *burnout* mahasiswa masih relatif terbatas, padahal indikator fisiologis dapat mencerminkan respons tubuh terhadap stres berkepanjangan.

Penelitian ini bertujuan untuk mengklasifikasikan tingkat burnout mahasiswa menggunakan algoritma *Random Forest* dengan dukungan teknik SMOTE untuk penyeimbangan data. Selain itu juga, untuk menganalisis *feature importance* guna mengidentifikasi faktor-faktor yang paling berpengaruh terhadap *burnout* mahasiswa. Dengan menggabungkan faktor psikologis, fisiologis, lingkungan, akademik, dan sosial, penelitian ini diharapkan dapat memberikan kontribusi tidak hanya dalam aspek performa klasifikasi, tetapi juga dalam penyediaan wawasan berbasis data yang dapat dimanfaatkan sebagai dasar pengembangan sistem deteksi dini *burnout* dan pengambilan keputusan di lingkungan perguruan tinggi, sebagaimana sistem pendukung keputusan telah terbukti efektif dalam membantu evaluasi kinerja sumber daya manusia di instansi lain [7].

2. Metodologi Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *machine learning* untuk mengklasifikasikan tingkat burnout mahasiswa. Algoritma utama yang digunakan adalah *Random Forest*, dengan analisis lanjutan berupa *feature importance* untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap *burnout*. Tahapan penelitian disusun secara sistematis mulai dari pengumpulan data hingga evaluasi kinerja model.

2.1. Alur Penelitian



Gbr1. Diagram Alur Penelitian

Alur penelitian dalam penelitian ini mengikuti tahapan umum penelitian berbasis *machine learning* yang meliputi pengumpulan data, pra-pemrosesan data, pembagian data, penyeimbangan data, pelatihan model, evaluasi model, dan analisis hasil. Pendekatan bertahap ini umum digunakan dalam penelitian klasifikasi data karena memastikan proses eksperimen yang sistematis dan replikatif, sebagaimana kerangka kerja yang diterapkan dalam studi pengolahan data terstruktur [8]. Menurut [9], alur penelitian yang terstruktur penting untuk menjamin konsistensi dan validitas hasil analisis dalam penemuan pengetahuan. Selain itu, analisis *feature importance* digunakan sebagai tahap lanjutan untuk meningkatkan interpretabilitas model *Random Forest*, mengingat algoritma ini memiliki kompleksitas struktur pohon yang tinggi yang memerlukan metode interpretasi khusus agar transparan [4].

2.2. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle *"Student Stress Monitoring Datasets"* tahun 2025, berupa dataset burnout mahasiswa dengan jumlah sampel 1.100 data yang terdiri dari 20 fitur dan 1 target label. Dataset tersebut memuat sejumlah fitur yang merepresentasikan kondisi mahasiswa dari berbagai aspek, meliputi faktor psikologis, fisiologis, lingkungan, akademik, dan sosial. Data dikumpulkan dalam bentuk data sekunder yang telah tersedia secara publik dan dipilih karena relevan dengan tujuan penelitian serta memiliki struktur yang sesuai untuk penerapan metode klasifikasi berbasis *machine learning*. Seluruh data yang digunakan pada penelitian ini bersifat anonim sehingga tidak memuat identitas pribadi responden.

2.3. Pra-Pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk memahami karakteristik dataset serta memastikan data siap digunakan dalam proses pemodelan. Seperti dijelaskan dalam studi pengolahan data, pra-pemrosesan merupakan langkah krusial yang meliputi pembersihan, integrasi, dan transformasi agar proses penemuan pola menjadi lebih mudah dan tepat [8]. Tahapan teknis yang dilakukan dalam penelitian ini mencakup pemeriksaan duplikasi data dan *missing values*, analisis korelasi antar fitur, serta seleksi fitur untuk mengurangi redundansi variabel yang berpotensi memengaruhi kinerja model [6]. Adapun langkah-langkah pra-pemrosesan data yang dilakukan adalah sebagai berikut:

1. *Data understanding*, untuk memahami struktur dataset, jumlah data, jenis variabel, serta distribusi data pada setiap fitur dan kelas burnout.
2. Pemeriksaan data duplikat dan *missing values*, untuk memastikan tidak terdapat data ganda maupun nilai kosong yang dapat memengaruhi proses analisis.
3. Analisis korelasi antar fitur, bertujuan untuk mengidentifikasi hubungan antar variabel dan menghindari redundansi fitur yang berpotensi menurunkan kinerja model.
4. Seleksi fitur atau variabel, dilakukan berdasarkan hasil analisis korelasi dan relevansi fitur terhadap variabel target, sehingga hanya fitur yang memiliki kontribusi penting yang digunakan dalam pemodelan.

2.4. Pembagian Data *(Data Splitting)*

Pembagian data dilakukan untuk memisahkan dataset menjadi data latih dan data uji sebelum proses penyeimbangan data dan pelatihan model. Tahap ini bertujuan untuk memastikan bahwa proses evaluasi model dilakukan secara objektif menggunakan data yang tidak terlibat dalam proses pelatihan. Dalam penelitian ini, pembagian data dilakukan menggunakan beberapa skenario perbandingan data latih dan data uji, yaitu 70:30, 75:25, 78:22, 80:20, 83:17, dan 87:13. Penggunaan beberapa skenario pembagian data bertujuan untuk menguji kestabilan dan konsistensi kinerja model *Random Forest* terhadap variasi proporsi data latih dan data uji, sebagaimana pendekatan eksperimental yang membandingkan performa model pada rasio pembagian data yang berbeda (seperti 70:30 dan 80:20) untuk mengevaluasi generalisasi model [10]. Selain itu, rasio 80:20 juga sering digunakan sebagai standar pembagian yang seimbang dalam klasifikasi stres mahasiswa untuk memastikan model memiliki data latih yang cukup [6][11].

Strategi ini digunakan sebagai alternatif terhadap *cross-validation* untuk menjaga kesederhanaan eksperimen serta menghindari peningkatan kompleksitas komputasi, mengingat fokus penelitian ini adalah analisis performa dan interpretabilitas model. Selain untuk menguji sensitivitas model terhadap variasi ukuran data latih, strategi ini juga bertujuan untuk mengamati konsistensi performa *Random Forest* pada distribusi data yang berbeda. Dengan membandingkan beberapa rasio pembagian data, penelitian ini dapat menunjukkan bahwa hasil klasifikasi tidak bergantung pada satu konfigurasi tertentu, melainkan stabil pada berbagai skenario. Hal ini memberikan gambaran yang lebih praktis terhadap penerapan model dalam kondisi nyata, di mana ketersediaan data latih dan data uji dapat bervariasi.

Metode *cross-validation* tidak digunakan secara eksplisit dalam penelitian ini karena algoritma *Random Forest* secara inheren telah memiliki mekanisme validasi internal melalui pendekatan *bootstrap aggregating* (*bagging*). Mekanisme ini bekerja dengan mengambil sampel data secara acak untuk membangun setiap pohon keputusan dan menggunakan data *out-of-bag* untuk validasi, yang secara efektif mengurangi varians dan mencegah *overfitting* tanpa memerlukan prosedur validasi silang eksternal yang kompleks [12]. Oleh karena itu, evaluasi menggunakan beberapa skenario pembagian data dinilai memadai untuk menunjukkan kestabilan model tanpa menambah beban komputasi secara signifikan.

2.5. Penyeimbangan Data menggunakan SMOTE

Ketidakseimbangan jumlah data antar kelas dapat menyebabkan penurunan kinerja model klasifikasi. Oleh karena itu, digunakan *Synthetic Minority Over-sampling Technique* (SMOTE) untuk meningkatkan representasi kelas minoritas pada data latih [11]. Erlin et al. [5] menunjukkan bahwa SMOTE mampu meningkatkan performa klasifikasi pada data tidak seimbang tanpa menghilangkan karakteristik data asli. Penerapan SMOTE hanya dilakukan pada data latih untuk mencegah terjadinya *data leakage*. Selain itu, penerapan SMOTE bertujuan untuk membantu model *Random Forest* dalam mempelajari pola data dari setiap kelas secara lebih seimbang, sehingga model tidak cenderung bias terhadap kelas mayoritas dan mampu meningkatkan kemampuan generalisasi pada data uji.

SMOTE bekerja dengan menghasilkan sampel sintetis baru pada kelas minoritas berdasarkan pendekatan *k-nearest neighbors*. Untuk setiap sampel pada kelas minoritas, algoritma terlebih dahulu mengidentifikasi sejumlah tetangga terdekat (biasanya ditentukan oleh parameter *k*) dalam ruang fitur. Selanjutnya, sampel sintetis dibangkitkan dengan melakukan interpolasi linear antara sampel asli dan salah satu tetangga terdekatnya. Proses ini dilakukan dengan memilih titik acak pada garis yang menghubungkan kedua sampel tersebut, sehingga data baru yang dihasilkan tetap berada dalam ruang distribusi kelas minoritas dan mempertahankan karakteristik statistiknya.

Berbeda dengan metode *random oversampling* yang hanya menggandakan data minoritas, SMOTE menciptakan variasi baru yang lebih representatif sehingga membantu memperluas wilayah keputusan (*decision region*) kelas minoritas. Dengan memperkaya distribusi data minoritas secara sintetis, batas pemisah antar kelas menjadi lebih jelas dan tidak terlalu dipengaruhi oleh dominasi kelas mayoritas. Pendekatan ini sangat penting dalam model berbasis pohon seperti Random Forest, karena distribusi data yang lebih seimbang memungkinkan proses pembentukan *decision tree* menghasilkan pemisahan node yang lebih optimal dan tidak bias terhadap kelas tertentu.

2.6. Pelatihan Model *Random Forest*

Tahap pelatihan model dilakukan menggunakan algoritma *Random Forest* dengan memanfaatkan data latih yang telah melalui proses pembagian data dan penyeimbangan menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). *Random Forest* merupakan metode *ensemble learning* yang membangun sejumlah pohon keputusan (*decision tree*) secara acak dan menghasilkan prediksi berdasarkan mekanisme agregasi (*majority voting*) dari seluruh pohon yang terbentuk. Menurut Karim et al. [13] dan Hu et al. [14], *Random Forest* memiliki keunggulan dalam menangani data berdimensi tinggi serta hubungan nonlinier antar variabel, sehingga banyak digunakan dalam permasalahan klasifikasi pada domain pendidikan dan kesehatan, sejalan dengan tren pengembangan sistem cerdas untuk pemantauan kondisi psikologis saat ini [3].

Pada penelitian ini, proses pelatihan model dilakukan secara terpisah untuk setiap skenario pembagian data. Model *Random Forest* dilatih menggunakan fitur-fitur terpilih yang merepresentasikan faktor psikologis, fisiologis, lingkungan, akademik, dan sosial. Pendekatan ini sejalan dengan rekomendasi Haddouchi dan Berrado [4], yang menyatakan bahwa *Random Forest* efektif digunakan pada data multivariat serta memungkinkan analisis lanjutan melalui *feature importance*. Model yang telah dilatih selanjutnya digunakan pada tahap evaluasi untuk mengukur kinerja klasifikasi burnout mahasiswa.

2.7. Evaluasi Model

Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

Evaluasi kinerja model dilakukan untuk menilai kemampuan algoritma *Random Forest* dalam mengklasifikasikan tingkat *burnout* mahasiswa secara menyeluruh. Penggunaan parameter pengujian yang terukur sangat penting dalam studi komparasi teknologi untuk memastikan metode yang dipilih memiliki performa yang optimal dan objektif [15]. Selain itu, penggunaan lebih dari satu metrik evaluasi diperlukan agar penilaian kinerja model tidak bias terhadap satu aspek tertentu, terutama pada dataset dengan distribusi kelas tidak seimbang. Visualisasi hasil klasifikasi dilakukan menggunakan *Confusion Matrix* untuk memetakan jumlah prediksi yang benar (*True Positive/True Negative*) dan salah (*False Positive/False Negative*), sehingga memudahkan analisis pola kesalahan klasifikasi [16]. Metrik evaluasi yang digunakan dalam penelitian ini merujuk pada standar pengukuran klasifikasi sebagai berikut [17][11]:

1. *Accuracy*, mengukur proporsi prediksi yang benar terhadap seluruh data uji dan memberikan gambaran umum tingkat ketepatan model.

$$Accuracy: (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

2. *Precision*, Mengukur proporsi prediksi positif yang benar terhadap seluruh prediksi positif yang dihasilkan oleh model, sehingga mencerminkan keandalan model dalam memprediksi kelas positif.

$$Precision: (TP) / (TP+FP) \quad (2)$$

3. *Recall*, Mengukur kemampuan model dalam mengidentifikasi seluruh data yang termasuk ke dalam kelas positif, yang penting pada kasus klasifikasi data tidak seimbang.

$$Recall: (TP) / (TP+FN) \quad (3)$$

4. *F1-Score*, Merupakan rata-rata harmonik antara *precision* dan *recall* yang digunakan untuk menyeimbangkan kedua metrik tersebut, khususnya ketika distribusi kelas tidak seimbang.

$$F1\text{-}Score: 2 \times (Precission \times Recall) / (Precision + Recall) \quad (3)$$

5. *Confusion Matrix*, untuk memvisualisasikan hasil klasifikasi model dalam bentuk matriks yang menunjukkan nilai *true positive* (TP), *false positive* (FP), *true negative* (TN), dan *false negative* (FN), sehingga memudahkan analisis pola kesalahan klasifikasi [18][16].

2.8. Analisa *Feature Importance*

Analisis *feature importance* dilakukan untuk mengidentifikasi kontribusi masing-masing variabel terhadap hasil klasifikasi *burnout* mahasiswa menggunakan algoritma *Random Forest*. Nilai *feature importance* diperoleh berdasarkan tingkat kontribusi setiap fitur dalam mengurangi ketidakmurnian (*impurity*) pada proses pembentukan pohon keputusan, sehingga mencerminkan pengaruh relatif fitur terhadap prediksi model [11]. Pendekatan ini digunakan untuk meningkatkan interpretabilitas model dan memahami hubungan antara variabel input dan tingkat *burnout* mahasiswa secara lebih mendalam. Menurut Haddouchi dan Berrado [4], analisis *feature importance* penting untuk menjembatani hasil komputasi dengan pengambilan keputusan berbasis data karena mampu membuka "kotak hitam" (*black box*) algoritma. Hasil analisis ini selanjutnya dimanfaatkan untuk mengidentifikasi faktor dominan yang berkontribusi terhadap *burnout* mahasiswa dan sebagai dasar pendukung pengembangan sistem deteksi dini seperti media konsultasi kesehatan mental otomatis [3] serta strategi pencegahan yang lebih tepat sasaran [12].

3. Hasil dan Pembahasan

Pada bagian ini disajikan hasil penelitian yang diperoleh dari seluruh tahapan eksperimen yang telah dilakukan, mulai dari pembagian dan penyeimbangan data, pelatihan model *Random Forest*, hingga evaluasi kinerja dan analisis *feature importance*. Penyajian hasil difokuskan pada informasi yang relevan dan mendukung tujuan penelitian, serta diuraikan secara sistematis untuk memberikan pemahaman yang jelas mengenai performa model dan faktor-faktor yang berpengaruh dalam klasifikasi burnout mahasiswa. Pembahasan dilakukan dengan mengaitkan hasil yang diperoleh dengan konteks penelitian sehingga dapat memberikan interpretasi yang bermakna dan aplikatif.

3.1. *Data Understanding*

Tahap *data understanding* dilakukan untuk memperoleh pemahaman awal mengenai karakteristik dan struktur dataset yang digunakan dalam penelitian. Dataset terdiri dari 1.100 data mahasiswa yang diperoleh dari platform Kaggle dan memuat sejumlah variabel yang merepresentasikan faktor fisiologis, psikologis, akademik, sosial, dan lingkungan. Pada tahap ini dilakukan identifikasi terhadap tipe data, kategori variabel, serta distribusi kelas target guna memastikan kesesuaian data dengan tujuan penelitian.

Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

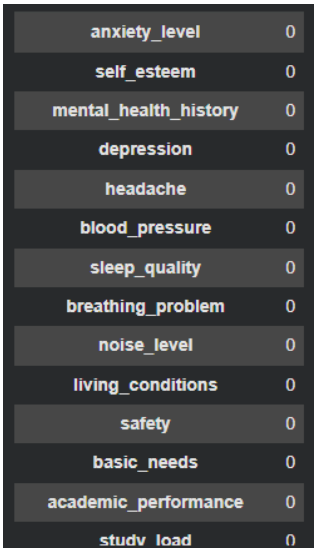
Tabel 1. Faktor, Fitur Utama, Tipe Data, Deskripsi

Faktor	Fitur Utama	Tipe Data	Deskripsi
Fisiologis	*Headache, Blood_Pressure, Sleep_Quality, Breathing_Problem*	Numerik	Kondisi Fisik
Psikologis	*Anxiety_Level, Self_Esteem, Mental_Health_History, Depression*	Numerik	Kondisi Mental
Akademik	*Academic_Performance, Study_Load, Teacher_Student_Relationship, Future_Career_Concerns*	Numerik	Aktivitas Akademik
Sosial	*Social_Support, Peer_Pressure, Extracurricular_Activities, Bullying*	Numerik	Interaksi Sosial
Lingkungan	*Noise_Level, Living_Conditions, Safety, Basic_Needs*	Numerik	Kondisi Lingkungan
Target	*Stress_Level* (3 Kelas)	Kategorikal	Tingkat Burnout Mahasiswa

Berdasarkan hasil eksplorasi awal, fitur-fitur dalam dataset mencerminkan pendekatan multidimensional terhadap *burnout* mahasiswa. Kombinasi faktor fisiologis, psikologis, akademik, sosial, dan lingkungan memberikan landasan yang komprehensif dalam membangun model klasifikasi. Tahap ini menjadi dasar penting sebelum dilakukan proses pra-pemrosesan lanjutan dan pemodelan, karena pemahaman struktur dan jenis data akan memengaruhi strategi analisis serta pendekatan klasifikasi yang digunakan.

3.2. Pemeriksaan *Missing Value* dan Duplikasi Data

Tahap pemeriksaan *missing value* dan duplikasi data dilakukan untuk memastikan kualitas dataset sebelum memasuki proses pemodelan. Data yang tidak lengkap atau terduplikasi dapat memengaruhi proses pelatihan model dan menghasilkan bias dalam evaluasi kinerja. Oleh karena itu, dilakukan pengecekan terhadap seluruh variabel pada dataset guna mengidentifikasi kemungkinan adanya nilai kosong (*null value*) maupun baris data yang terduplikasi.



Gbr *2*. Hasil Pemeriksaan *Missing Value*

```
df.duplicated().sum()
np.int64(0)
```

Gbr *3*. Hasil Pemeriksaan Duplikasi Data

Berdasarkan hasil pemeriksaan, tidak ditemukan nilai kosong maupun data duplikat dalam dataset yang digunakan. Hal ini menunjukkan bahwa dataset berada dalam kondisi yang bersih dan siap digunakan untuk tahap analisis lanjutan tanpa memerlukan proses imputasi atau penghapusan data. Dengan kualitas data yang terjaga, proses pelatihan model dapat dilakukan secara lebih optimal dan hasil evaluasi yang diperoleh menjadi lebih representatif terhadap kondisi sebenarnya.

3.3. Analisis Korelasi Fitur

Analisis korelasi fitur dilakukan untuk mengidentifikasi tingkat hubungan antar variabel dalam dataset sebelum proses pemodelan dilakukan. Berdasarkan hasil visualisasi matriks korelasi, sebagian besar fitur menunjukkan hubungan yang berada pada kategori sedang, dengan nilai korelasi berkisar antara ±0,3 hingga ±0,7. Hal ini menunjukkan adanya keterkaitan antar beberapa variabel, namun tidak pada tingkat yang ekstrem sehingga masih dapat digunakan secara bersamaan dalam model klasifikasi.

Beberapa pasangan variabel menunjukkan korelasi yang relatif kuat. Sebagai contoh, *anxiety_level* memiliki korelasi positif yang tinggi dengan *future_career_concerns* (sekitar 0,72) dan *bullying* (sekitar 0,71), serta korelasi negatif yang cukup kuat dengan *sleep_quality* (sekitar -0,71). Selain itu, *depression* juga menunjukkan hubungan positif yang kuat dengan *future_career_concerns* dan *bullying*. Hubungan-hubungan ini secara konseptual konsisten, karena tekanan psikologis dan kekhawatiran terhadap masa depan dapat berkaitan erat dengan tingkat kecemasan dan depresi mahasiswa.

Meskipun terdapat beberapa korelasi yang cukup tinggi, tidak ditemukan nilai korelasi yang mendekati ±1 yang mengindikasikan redundansi sempurna antar fitur. Dengan demikian, tidak terdapat indikasi multikolinearitas yang ekstrem yang dapat mengganggu proses pelatihan model *Random Forest*. Mengingat *Random Forest* merupakan algoritma berbasis pohon yang relatif *robust* terhadap korelasi antar variabel, seluruh fitur tetap dipertahankan dalam proses pemodelan untuk menjaga kelengkapan informasi dalam klasifikasi *burnout* mahasiswa.

3.4. Seleksi Fitur

Tahap seleksi fitur dilakukan untuk mengurangi redundansi informasi dan mempertahankan variabel yang paling relevan dalam proses klasifikasi *burnout* mahasiswa. Berdasarkan hasil analisis korelasi serta pertimbangan kontribusi variabel

Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

terhadap target, jumlah fitur awal sebanyak 20 variabel dikurangi menjadi 15 variabel. Proses ini bertujuan untuk meningkatkan efisiensi model serta mengurangi potensi gangguan akibat fitur yang memiliki korelasi tinggi atau kontribusi yang relatif rendah terhadap proses klasifikasi.

Lima variabel yang dieliminasi pada tahap seleksi adalah *peer_pressure*, *living_conditions*, *mental_health_history*, *sleep_quality*, dan *future_career_concerns*. Penghapusan variabel tersebut dilakukan karena menunjukkan korelasi yang cukup kuat dengan variabel lain atau memiliki representasi informasi yang tumpang tindih dengan fitur psikologis maupun sosial lainnya. Dengan mengurangi fitur yang bersifat redundan, model diharapkan dapat mempelajari pola data secara lebih optimal tanpa kehilangan informasi yang signifikan.

Setelah proses seleksi, sebanyak 15 fitur dipertahankan, yaitu *anxiety_level*, *self_esteem*, *depression*, *headache*, *blood_pressure*, *breathing_problem*, *noise_level*, *safety*, *basic_needs*,*academic_performance*,*study_load*,*teacher_student_relationship*,*social_support*,*extracurricular_activities*,*bullying*, dan *stress_level*. Fitur-fitur tersebut dinilai mewakili dimensi fisiologis, psikologis, akademik, sosial, dan lingkungan secara komprehensif. Dengan demikian, dataset hasil seleksi menjadi lebih ringkas namun tetap informatif untuk mendukung proses pelatihan model *Random Forest* pada tahap berikutnya.

3.5. Pembagian Data *(Data Splitting)*

Pada tahap ini, dataset *burnout* mahasiswa yang berjumlah 1.100 data dibagi menjadi data latih dan data uji menggunakan enam skenario pembagian data, yaitu 70:30, 75:25, 78:22, 80:20, 83:17, dan 87:13. Pembagian data dilakukan untuk mengevaluasi pengaruh variasi proporsi data latih dan data uji terhadap kinerja model *Random Forest*.

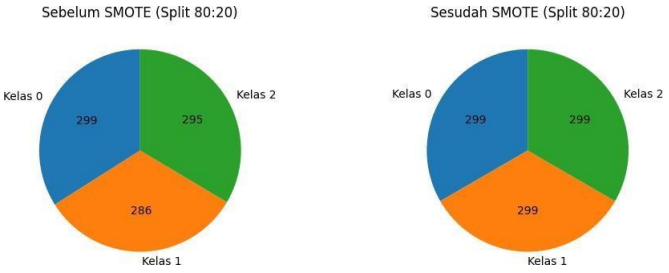
Tabel 2. *Splitting* Data

Split	Training (%)	Testing (%)	Jumlah Data Training	Jumlah Data Testing
70:30	70	30	770	330
75:25	75	25	825	275
78:22	78	22	858	242
80:20	80	20	880	220
83:17	83	17	913	187
87:13	87	13	957	143

Hasil pembagian data menunjukkan bahwa setiap skenario menghasilkan jumlah data latih dan data uji yang berbeda, sebagaimana ditunjukkan pada Tabel 1. Semakin besar proporsi data latih, semakin banyak data yang digunakan untuk proses pembelajaran model, sementara jumlah data uji semakin berkurang. Pembagian data ini menjadi dasar bagi tahapan selanjutnya, yaitu penyeimbangan data menggunakan SMOTE dan pelatihan model klasifikasi.

3.6. Penyeimbangan Data menggunakan SMOTE

Penyeimbangan data dilakukan menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) dan diterapkan hanya pada data latih dari setiap skenario pembagian data. Penerapan SMOTE bertujuan untuk mengatasi ketidakseimbangan jumlah data antar kelas *burnout* tanpa memengaruhi data uji, sehingga proses evaluasi model tetap objektif. Pada penelitian ini, penyeimbangan data difokuskan pada skenario pembagian data 80:20 karena skenario tersebut menghasilkan performa model terbaik pada tahap evaluasi.



Gbr 4. Distribusi Kelas Data Latih Sebelum dan Sesudah SMOTE

Hasil penyeimbangan menunjukkan bahwa distribusi kelas pada data latih menjadi sepenuhnya seimbang setelah penerapan SMOTE. Sebelum penyeimbangan, jumlah data pada masing-masing kelas masih menunjukkan perbedaan, sedangkan setelah SMOTE diterapkan jumlah data pada setiap kelas menjadi sama, yaitu sebanyak 299 data. Kondisi ini memungkinkan model *Random Forest* mempelajari pola data dari setiap kelas secara lebih merata serta mengurangi potensi bias terhadap kelas tertentu pada proses pelatihan model.

3.7. Pelatihan dan Evaluasi Model *Random Forest*

Pada tahap ini, model *Random Forest* dilatih menggunakan data latih yang telah diseimbangkan dengan SMOTE pada setiap skenario pembagian data. Evaluasi kinerja model dilakukan menggunakan data uji dengan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil evaluasi difokuskan pada perbandingan performa model antar skenario pembagian data untuk mengetahui pengaruh proporsi data latih dan data uji terhadap kinerja klasifikasi *burnout* mahasiswa.

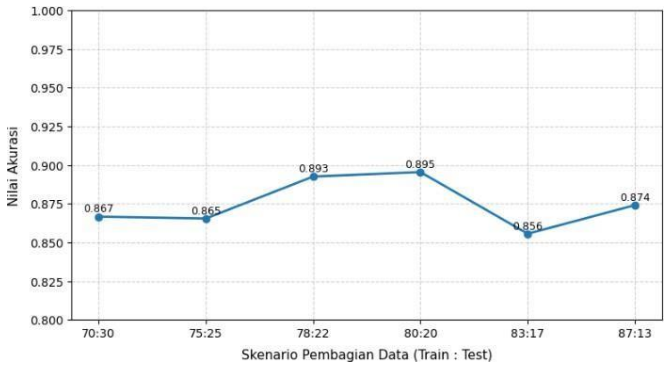
Tabel 3. Hasil *Classification Report* Setiap *Splitting* Data

Split	*Accuracy (%)*	*Precision*	*Recall*	*F1-Score*
70:30	86.67	0,87	0.87	0.87
75:25	86.55	0.87	0.87	0.87
78:22	89.26	0.89	0.89	0.89
80:20	89.55	0.90	0.90	0.90
83:17	85.56	0.86	0.86	0.86
87:13	87.41	0.88	0.87	0.87

Ringkasan hasil akurasi model *Random Forest* untuk seluruh skenario pembagian data disajikan pada Tabel 2. Berdasarkan hasil tersebut, seluruh skenario menghasilkan nilai akurasi di atas 85%, yang menunjukkan bahwa model

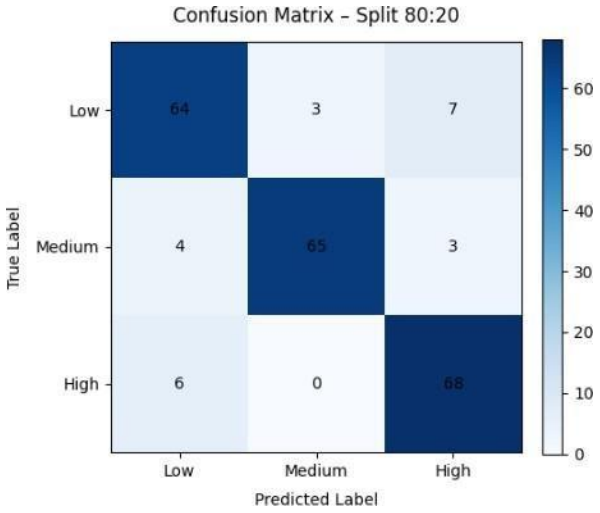
Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

mampu melakukan klasifikasi burnout mahasiswa dengan performa yang baik dan relatif stabil. Skenario pembagian data 80:20 menghasilkan nilai akurasi tertinggi sebesar 89,55%, diikuti oleh skenario 78:22 dengan akurasi 89,26%.



Gbr 5. Grafik Perbandingan Akurasi antar Skenario Splitting Data

Perbandingan akurasi antar skenario splitting data divisualisasikan pada Gbr 5, yang menunjukkan bahwa perubahan proporsi data latih dan data uji tidak menyebabkan fluktuasi performa yang signifikan. Stabilitas ini mengindikasikan bahwa model *Random Forest* memiliki kemampuan generalisasi yang baik setelah penerapan penyeimbangan data menggunakan SMOTE. Dengan demikian, kombinasi *Random Forest* dan SMOTE terbukti efektif dalam mengklasifikasikan tingkat *burnout* mahasiswa pada berbagai skenario pembagian data.

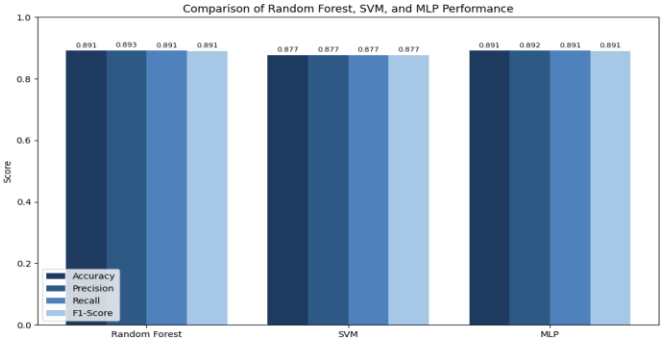


Gbr 6. Confusion Matrix

Perbandingan nilai akurasi antar skenario divisualisasikan pada Gbr 5 dan menunjukkan bahwa perubahan proporsi data latih dan data uji tidak menyebabkan fluktuasi performa yang signifikan. Untuk memberikan gambaran yang lebih rinci mengenai kinerja model pada skenario terbaik, ditampilkan *confusion matrix* untuk skenario pembagian data 80:20 pada Gbr 6. Hasil *confusion matrix* menunjukkan bahwa sebagian besar data pada masing-masing kelas *burnout* berhasil diklasifikasikan dengan benar, dengan jumlah kesalahan klasifikasi yang relatif kecil, sehingga memperkuat hasil evaluasi yang diperoleh dari metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

3.8. Perbandingan Kinerja *Random Forest, SVM,* dan *MLP*

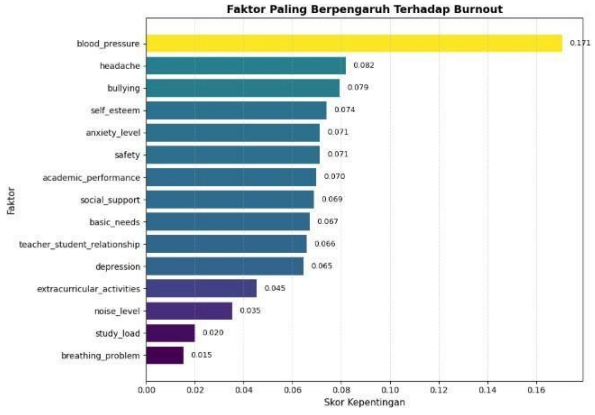
Untuk memastikan robustness model yang digunakan, dilakukan perbandingan kinerja antara *Random Forest*, *Support Vector Machine* (SVM), dan *Multilayer Perceptron* (MLP) pada skenario pembagian data terbaik. Perbandingan dilakukan menggunakan metrik evaluasi *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk memberikan gambaran menyeluruh terhadap performa klasifikasi masing-masing algoritma.



Gbr 7. Perbandingan kinerja *Random Forest*, SVM, dan MLP

3.9. Analisis *Feature Importance*

Analisis *feature importance* dilakukan untuk mengidentifikasi faktor-faktor yang paling berpengaruh dalam proses klasifikasi *burnout* mahasiswa menggunakan model *Random Forest*. Hasil analisis yang ditampilkan pada Gbr 8 menunjukkan bahwa faktor fisiologis *blood_pressure* memiliki skor kepentingan tertinggi dibandingkan faktor lainnya, sehingga menjadi variabel paling dominan dalam menentukan tingkat *burnout* mahasiswa.



Jurnal Pseudocode Vol 13 (1) Februari 2026, hal 54-62

Gbr 8. Grafik *Feature Importance*

Selain *blood_pressure*, beberapa faktor lain seperti *headache*, *bullying*, *self_esteem*, *anxiety_level*, dan *academic_performance* juga menunjukkan kontribusi yang cukup signifikan, meskipun nilainya berada di bawah faktor fisiologis. Sementara itu, faktor seperti *study_load*, *noise_level*, dan *breathing_problem* memiliki skor kepentingan yang relatif lebih rendah. Temuan ini menunjukkan bahwa burnout mahasiswa dipengaruhi oleh kombinasi faktor fisiologis, psikologis, akademik, dan sosial, dengan faktor fisiologis sebagai kontributor utama dalam model klasifikasi yang digunakan.

3.10. Pembahasan

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* mampu memberikan kinerja klasifikasi yang baik dan stabil dalam mengidentifikasi tingkat *burnout* mahasiswa. Hal ini tercermin dari nilai akurasi yang relatif konsisten pada seluruh skenario pembagian data yang diuji. Stabilitas performa ini menunjukkan bahwa model memiliki kemampuan yang baik dalam menangkap pola hubungan antar variabel yang kompleks, terutama pada data dengan karakteristik multidimensi yang melibatkan faktor psikologis, fisiologis, lingkungan, akademik, dan sosial.

Beberapa skenario pembagian data memberikan gambaran yang lebih komprehensif mengenai pengaruh proporsi data latih dan data uji terhadap kinerja model. Dari enam skenario yang diuji, pembagian data 80:20 menghasilkan performa terbaik. Kondisi ini menunjukkan bahwa jumlah data latih yang lebih besar memungkinkan model mempelajari karakteristik data secara lebih optimal, sementara jumlah data uji masih mencukupi untuk mengukur kemampuan generalisasi model secara objektif. Temuan ini menegaskan pentingnya pemilihan proporsi data yang tepat dalam proses pemodelan klasifikasi.

Penerapan penyeimbangan data menggunakan SMOTE yang dilakukan hanya pada data latih terbukti menjadi langkah yang tepat dalam mengatasi ketidakseimbangan kelas. Dengan distribusi kelas yang lebih seimbang, model *Random Forest* memiliki kesempatan yang sama untuk mempelajari pola dari setiap kelas *burnout*. Hal ini berkontribusi pada peningkatan kualitas pembelajaran model serta membantu mengurangi potensi bias terhadap kelas tertentu, sebagaimana tercermin dari hasil evaluasi dan *confusion matrix* pada skenario terbaik.

Hasil *confusion matrix* pada skenario pembagian data 80:20 menunjukkan bahwa sebagian besar data pada setiap kelas burnout berhasil diklasifikasikan dengan benar, dengan jumlah kesalahan klasifikasi yang relatif kecil. Temuan ini memperlihatkan bahwa model tidak hanya memiliki akurasi yang tinggi secara keseluruhan, tetapi juga mampu membedakan antar kelas *burnout* dengan cukup baik. Dengan demikian, model yang dihasilkan memiliki potensi untuk digunakan dalam konteks nyata yang memerlukan pemetaan tingkat *burnout* secara lebih akurat.

Selain evaluasi kinerja model, analisis *feature importance* memberikan kontribusi penting dalam meningkatkan interpretabilitas hasil klasifikasi. Dominannya variabel *blood_pressure* menunjukkan bahwa faktor fisiologis memiliki peran yang signifikan dalam menentukan tingkat burnout mahasiswa. Temuan ini memperkuat pandangan bahwa *burnout* bukan hanya fenomena psikologis atau akademik semata, tetapi juga berkaitan erat dengan respons fisiologis individu terhadap tekanan yang dialami secara berkelanjutan.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa pendekatan klasifikasi berbasis *Random Forest* yang dikombinasikan dengan penyeimbangan data dan analisis *feature importance* mampu memberikan hasil yang tidak hanya akurat, tetapi juga mudah diinterpretasikan. Integrasi antara kinerja model dan pemahaman terhadap faktor dominan burnout memberikan dasar yang kuat bagi pemanfaatan pendekatan berbasis data dalam memahami dan mendeteksi burnout mahasiswa secara lebih komprehensif di lingkungan perguruan tinggi.

4. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa pendekatan klasifikasi berbasis *machine learning* dapat digunakan secara efektif untuk memodelkan dan memahami fenomena *burnout* mahasiswa yang bersifat multidimensional. Hasil eksperimen memperlihatkan bahwa algoritma *Random Forest* mampu memberikan kinerja yang stabil pada berbagai skenario pembagian data, yang menandakan bahwa model memiliki kemampuan generalisasi yang baik terhadap variasi distribusi data. Temuan ini menegaskan bahwa pemilihan algoritma yang tepat, dikombinasikan dengan alur pemrosesan data yang benar, berperan penting dalam menghasilkan model klasifikasi yang andal.

Penerapan penyeimbangan data menggunakan SMOTE yang dilakukan secara terbatas pada data latih terbukti menjadi langkah krusial dalam menjaga objektivitas evaluasi dan meningkatkan kualitas pembelajaran model. Pendekatan ini memungkinkan model mempelajari karakteristik setiap kelas secara lebih merata tanpa menimbulkan *data leakage*. Dengan demikian, penelitian ini menekankan pentingnya penanganan ketidakseimbangan data sebagai bagian integral dari perancangan sistem klasifikasi, khususnya pada permasalahan yang melibatkan data sosial dan psikologis.

Analisis *feature importance* memberikan kontribusi penting dalam meningkatkan interpretabilitas model. Dominannya faktor fisiologis, khususnya *blood_pressure*, menunjukkan bahwa indikator fisik memiliki peran signifikan dalam membedakan tingkat *burnout* mahasiswa. Temuan ini memperluas pemahaman bahwa *burnout* tidak hanya merupakan persoalan akademik atau psikologis, tetapi juga berkaitan erat dengan respons fisiologis individu terhadap tekanan yang dialami secara berkelanjutan. Dengan adanya interpretasi fitur, hasil penelitian menjadi lebih bermakna dan tidak terbatas pada capaian performa numerik semata.

Secara keseluruhan, penelitian ini memiliki implikasi praktis dalam pengembangan sistem pendukung keputusan atau sistem deteksi dini burnout mahasiswa berbasis data. Model dan pendekatan yang digunakan dapat diterapkan sebagai dasar untuk pemantauan kondisi mahasiswa secara lebih komprehensif di lingkungan perguruan tinggi, serta dapat diperluas dengan integrasi data tambahan atau

pendekatan analitik lanjutan guna meningkatkan ketepatan dan cakupan analisis *burnout* di masa mendatang.

Referensi

[1] L. David, R. Wardhana, N. H. Yoenanto, N. A. Fardhana, F. Psikologi, and U. Airlangga, "Factors Affecting Academic Burnout in College Students : Scoping Review Faktor-Faktor yang Mempengaruhi Academic Burnout Pada Mahasiswa : Scoping Review," vol. 13, no. 2, pp. 201–210, 2025. http://dx.doi.org/10.30872/psikoborneo.v13i2.18351

[2] I. Damayanti and F. Aulia, "Adaptasi Academic Burnout Scale Versi Indonesia," vol. 6, no. 6, pp. 4653–4662, 2025. https://doi.org/10.38035/jmpis.v6i6

[3] A. Hadi, "Design of NLP-based chatbot media as a mental health consultation media with Depression Anxiety Stress approach," vol. 05, 2025. https://doi.org/10.31763/iota.v5i1.888

[4] M. Haddouchi and A. Berrado, "A survey and taxonomy of methods interpreting random forest models".

[5] Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang Impact of SMOTE on Random Forest Classifier Performance based on Imbalanced Data," vol. 21, no. 3, 2022, doi: 10.30812/matrik.v21i3.1726.

[6] V. Oktaviani, N. Rosmawarni, and M. P. Muslim, "Perbandingan Kinerja Random Forest Dan Smote Random Forest Dalam Mendeteksi Dan Mengukur Tingkat Stres Pada Mahasiswa Tingkat Akhir," vol. 4221, no. April, pp. 43–49, 2024.

[7] D. A. Prameswari, A. Hadi, I. Teknologi, D. Bisnis, and A. Malang, "Sistem Pendukung Keputusan Penilaian Kinerja Karyawan Pada Diskominfo Di Kabupaten Nganjuk Berbasis Web," vol. 17, no. 2, pp. 147–156, 2023. https://doi.org/10.32815/jitika.v17i2.931

[8] S. A. Hadi Malang, "Segmentasi Pelanggan Internet Service Provider (ISP) Berbasis Pillar K-Means," vol. 13, no. 2, pp. 151–159, 2019. https://doi.org/10.32815/jitika.v13i2.413

[9] X. Shu and Y. Ye, "Knowledge Discovery : Methods from data mining and machine learning ☆," *Soc. Sci. Res.*, vol. 110, no. October 2022, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.https://doi.org/10.1016/j.ssresearch.2022.102817

[10] A. A. Hapsari, A. S. Nursuwanda, H. Zuhriyah, and D. J. Vresdian, "Klasifikasi Kesehatan Mental Mahasiswa Model TMAS dengan Algoritma Decision Tree , Logistic Regression , dan Random Forest," vol. 7, no. November, 2024.

[11] Santoso, R. N. L. D. E., & Kusnawi. (2025). Optimization of Stress Classification Among Students Using Random Forest Algorithm. *SITEKNIK: Jurnal Sistem Informasi, Teknik dan Teknologi Terapan*, *2*(2), 76–87. https://doi.org/10.5281/zenodo.15130385

[12] A. Adinegara and S. Widodo, "Prediction of Academic Burnout in College Students : A Comparative Analysis of Support Vector Machine and Random Forest Algorithms Prediksi Academic Burnout pada Mahasiswa : Analisis Komparatif Algoritma Support Vector Machine dan Random Forest," vol. 5, no. October, pp. 1367–1376, 2025. https://doi.org/10.57152/malcom.v5i4.2283

[13] F. Karim, A. Oyewande, L. F. Abdalla, and R. C. Ehsanullah, "Social Media Use and Its Connection to Mental Health : A Systematic Review," vol. 12, no. 6, 2020, doi: 10.7759/cureus.8627. https://doi.org/10.7759/cureus.8627

[14] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random," vol. 24, no. January, pp. 1–11, 2023. https://doi.org/10.1093/bib/bbad002

[15] Zaqi, A., & Hadi, A. (2023). Komparasi Load Balancing Metode PCC dan NTH Pada Mikrotik Implentasi di Al Irsyad Tengaran 7 Batu. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *8*(1), 021–026. https://doi.org/10.29100/jipi.v8i1.3261

[16] N. L. Chusna, N. Wiliani, and A. F. Abdillah, "Enhancing Ulos Batik Pattern Recognition through Machine Learning : A Study with KNN and SVM," vol. 6, no. 3, 2024. https://doi.org/10.34288/jri.v6i3.311

[17] N. Kadek, W. Patrianingsih, K. Setemen, D. Gede, and H. Divayana, "Comparison Analysis Of Naïve Bayes And K - Nearest Neighbor Methods On The Prediction Of Academic Potential In Smk Ti Bali Global Badung," vol. 5, no. 36, pp. 1557–1566, 2021. www.iocscience.org/ejournal/index.php/mantik/index

[18] Nasrullah, A. H**.** (2021). Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris. *Jurnal Ilmiah Ilmu Komputer*, *5*(1), 45-53.