

Prediksi Partisipasi Mahasiswa dalam Kegiatan Kampus Menggunakan Algoritma Random Forest

Dika Setyawan^{1*}, Muhammad Faruq Ghullam Al Khoiri², Ivan Noviandra³, Andriyan Susetyo⁴, Nisa Aulia Ramadhani⁵, Satria Setyaji⁶

¹Teknik Informatika/Fakultas Ilmu Komputer

Universitas Duta Bangsa

^{1*}240103129@mhs.udb.ac.id

²Teknik Informatika/Fakultas Ilmu Komputer

Universitas Duta Bangsa

²240103139@mhs.udb.ac.id

³Teknik Informatika/Fakultas Ilmu Komputer

Universitas Duta Bangsa

³240103134@mhs.udb.ac.id

⁴Teknik Informatika/Fakultas Ilmu Komputer

Universitas Duta Bangsa

⁴240103125@mhs.udb.ac.id

⁵Teknik Informatika/Fakultas Ilmu Komputer

Universitas Duta Bangsa

⁵240103143@mhs.udb.ac.id

⁶Teknik Informatika/Fakultas Ilmu Komputer

Universitas Duta Bangsa

⁶240103257@mhs.udb.ac.id

Abstrak— Partisipasi mahasiswa dalam kegiatan kampus merupakan faktor penting yang memengaruhi keberhasilan akademik dan pengembangan kompetensi sosial. Namun, rendahnya tingkat keterlibatan mahasiswa masih menjadi tantangan di banyak perguruan tinggi Indonesia. Penelitian ini bertujuan membangun model prediksi partisipasi mahasiswa dalam kegiatan kampus menggunakan algoritma Random Forest berbasis data kuesioner dari 500 responden, sekaligus mengidentifikasi faktor-faktor dominan yang memengaruhinya. Data dikumpulkan melalui kuesioner terstruktur yang mencakup variabel motivasi, pengaruh teman sebaya, ketersediaan waktu, dan keterlibatan organisasi. Preprocessing data meliputi pembersihan data, encoding variabel kategorikal, dan pembagian data latih-uji. Model Random Forest dibangun dengan mekanisme ensemble learning melalui voting mayoritas dari sejumlah pohon keputusan. Hasil eksperimen menunjukkan bahwa model mencapai akurasi sebesar 74%, precision 66,67%, recall 42,42%, dan F1-score 51,84% dalam memprediksi tingkat partisipasi mahasiswa. Analisis feature importance mengungkapkan bahwa ketersediaan waktu luang (36%) dan status keikutsertaan organisasi saat ini (26,7%) merupakan faktor paling dominan dalam menentukan keputusan mahasiswa untuk berpartisipasi, sedangkan pengaruh ajakan teman memberikan kontribusi paling kecil (16,5%). Temuan ini diharapkan dapat menjadi dasar bagi institusi perguruan tinggi dalam merancang strategi intervensi yang lebih tepat sasaran guna meningkatkan keterlibatan mahasiswa secara berkelanjutan.

Kata kunci— partisipasi mahasiswa, kegiatan kampus, Random Forest, Educational Data Mining, machine learning

Abstract— Student participation in campus activities is a crucial factor influencing academic achievement and social competency development. However, low levels of student engagement remain a persistent challenge in many Indonesian higher education institutions. This study aims to develop a predictive model for student participation in campus activities using the Random Forest algorithm based on questionnaire data collected from 500 respondents, while simultaneously identifying the dominant factors that influence such participation. Data were gathered through a structured questionnaire covering variables including motivation, peer influence, time availability, and organizational involvement. Data preprocessing involved data cleaning, categorical variable encoding, and train-test splitting. The Random Forest model was constructed through an ensemble learning mechanism using majority voting across multiple decision trees. Experimental results demonstrate that the model achieves an accuracy of 74%, precision of 66.67%, recall of 42.42%, and F1-score of 51.84% in predicting student participation levels. Feature importance analysis reveals that time availability (36%) and current organizational involvement (26.7%) are the most dominant factors in determining students' decisions to participate, while peer influence contributes the least (16.5%). These findings are expected to serve as a foundation for higher education institutions in designing more targeted intervention strategies to sustainably improve student engagement.

Keywords— student participation, campus activities, Random Forest, Educational Data Mining, machine learning

I. PENDAHULUAN

Perguruan tinggi tidak hanya berperan sebagai pusat transfer ilmu pengetahuan, tetapi juga sebagai wadah pengembangan diri

mahasiswa melalui berbagai kegiatan kampus.

Keterlibatan aktif mahasiswa dalam kegiatan kampus terbukti berkontribusi positif terhadap

keberhasilan akademik dan pengembangan kompetensi sosial [1]. Namun, rendahnya tingkat partisipasi mahasiswa dalam kegiatan kampus masih menjadi tantangan nyata yang dihadapi oleh banyak perguruan tinggi di Indonesia [2]. Kondisi ini mendorong perlunya upaya sistematis untuk memahami dan memprediksi perilaku partisipasi mahasiswa secara lebih mendalam.

Berbagai faktor memengaruhi keputusan mahasiswa untuk berpartisipasi dalam kegiatan kampus, mulai dari ketersediaan waktu, motivasi diri, hingga pengaruh lingkungan sosial dan teman sebaya [3]. Tanpa pemahaman yang komprehensif terhadap faktor-faktor tersebut, upaya peningkatan partisipasi sulit dilakukan secara tepat sasaran oleh pihak institusi [4]. Dengan demikian, dibutuhkan metode analisis yang berlandaskan pada data yang mampu mengidentifikasi pola perilaku mahasiswa secara otomatis dan akurat.

Pendekatan Educational Data Mining (EDM) berbasis machine learning menjadi solusi yang relevan untuk menjawab tantangan tersebut [5]. Berbagai algoritma machine learning telah diterapkan dalam konteks pendidikan, salah satunya Random Forest. Algoritma ini merupakan teknik ensemble learning yang mengonstruksi sejumlah pohon keputusan secara bersamaan dan menghasilkan prediksi melalui mekanisme pemungutan suara terbanyak [6]. Penelitian sebelumnya menunjukkan bahwa Random Forest mampu mengungguli algoritma lain seperti Naive Bayes dan Decision Tree dalam prediksi berbasis data akademik mahasiswa [7]. Meskipun Random Forest tergolong handal untuk tugas klasifikasi, hasil akhir prediksinya tetap sangat bergantung pada mutu data yang digunakan. Oleh karena itu, tahap preprocessing dibutuhkan guna mempersiapkan data supaya layak digunakan dalam proses pelatihan model [8]. Setelah itu, performa model diukur memakai confusion matrix yang

mencakup empat indikator utama yakni accuracy, precision, recall, dan F1-score guna menilai kemampuan klasifikasinya [9].

Meskipun penerapan Random Forest pada konteks pendidikan bukan hal baru, sebagian besar penelitian terdahulu masih berfokus pada prediksi kinerja akademik atau kelulusan mahasiswa, sementara aspek partisipasi mahasiswa dalam kegiatan non-akademik seperti kegiatan kampus relatif belum banyak dikaji secara kuantitatif. Selain itu, studi-studi sebelumnya umumnya belum mengombinasikan variabel motivasi diri, pengaruh teman sebaya, ketersediaan waktu, dan keterlibatan organisasi secara bersamaan dalam satu model prediksi. Studi Jawad dkk. [1], Saputra [2], dan Lu dkk. [4] menerapkan Random Forest untuk memprediksi kinerja akademik atau risiko dropout mahasiswa. Balabied dan Eid [6] serta Wantos dan Bahri [8] turut memakai variabel akademik seperti nilai dan kehadiran kuliah sebagai prediktor utama. Penelitian ini berbeda pada dua sisi, variabel input berasal dari faktor perilaku dan sosial, bukan akademik, dan target prediksi adalah partisipasi kegiatan kampus, bukan performa akademik. Pemilihan Random Forest pada penelitian ini didasarkan pada temuan Fatunnisa dan Marcos [7] serta Hidayat dkk. [9], yang menunjukkan Random Forest mengungguli Decision Tree tunggal pada data pendidikan berskala kecil-menengah karena mekanisme bagging menurunkan varians dan mengurangi risiko overfitting. Karakteristik ini relevan dengan penelitian ini, mengingat ukuran sampel yang digunakan (500 responden) dan jumlah fitur (8 variabel) tergolong terbatas. Kesenjangan inilah yang menjadi dasar penelitian ini, yaitu mengisi celah tersebut dengan membangun model prediksi partisipasi mahasiswa dalam kegiatan kampus berbasis data kuesioner primer, sekaligus mengidentifikasi faktor-faktor yang paling dominan

memengaruhinya melalui analisis feature importance.

Berdasarkan uraian tersebut, Studi ini dimaksudkan untuk mengimplementasikan algoritma Random Forest dalam rangka membangun sebuah model klasifikasi dan mengevaluasi kinerjanya pada data yang digunakan. Hasil penelitian diharapkan dapat menjadi referensi dalam penerapan algoritma Random Forest pada penelitian sejenis [10].

II. METODOLOGI PENELITIAN

A. Pengumpulan Data

Data penelitian diperoleh melalui angket berbasis daring menggunakan Google Form yang didistribusikan kepada mahasiswa aktif Universitas Duta Bangsa Surakarta dan mahasiswa universitas lain. Proses pengumpulan data dilakukan selama kurang lebih 2 minggu dan berhasil mengumpulkan sebanyak 500 responden. Kuesioner dirancang untuk memperoleh informasi terkait faktor-faktor yang diduga memengaruhi partisipasi mahasiswa dalam kegiatan kampus. Variabel input yang dikumpulkan meliputi: (1) Jenis Kelamin, (2) Program Studi, (3) Semester Saat Ini, (4) Mengikuti Organisasi Saat Ini, (5) Pernah Mengikuti Organisasi Sebelumnya, (6) Pengaruh Ajakan Teman, (7) Memiliki Waktu Luang, dan (8) Kehadiran pada Kegiatan Kampus sebagai variabel target dengan nilai Hadir (1) dan Tidak Hadir (0). Teknik pengambilan sampel yang digunakan adalah convenience sampling yang dikombinasikan dengan penyebaran berantai (snowball) melalui aplikasi WhatsApp dan media sosial, mengingat keterbatasan akses terhadap kerangka sampel populasi mahasiswa secara menyeluruh. Kuesioner disebarluaskan secara langsung kepada responden tanpa melalui tahap pengujian validitas dan reliabilitas secara statistik (seperti expert judgment, korelasi item-total, maupun Cronbach's Alpha) terlebih dahulu,

sehingga ketiadaan uji tersebut menjadi salah satu keterbatasan penelitian ini yang perlu menjadi perhatian pada penelitian selanjutnya. Pemilihan kedelapan variabel input didasarkan pada tinjauan pustaka terkait faktor-faktor yang secara teoretis memengaruhi partisipasi mahasiswa, yaitu karakteristik demografis (jenis kelamin, program studi, semester), riwayat dan status keterlibatan organisasi, pengaruh sosial dari teman sebaya, serta ketersediaan waktu luang, sebagaimana diuraikan pada bagian Pendahuluan. Variabel Jenis Kelamin, Program Studi, dan Semester merujuk pada kajian karakteristik demografis dalam Lu dkk. [4], variabel Mengikuti Organisasi Saat Ini dan Pernah Mengikuti Organisasi Sebelumnya merujuk pada Putra dan Baskoro [3], sedangkan variabel Pengaruh Ajakan Teman dan Memiliki Waktu Luang merujuk pada kajian faktor sosial dan manajemen waktu mahasiswa dalam [3] dan [4].

B. Preprocessing Data

Tahap preprocessing dilaksanakan guna mengondisikan data sehingga siap digunakan dalam proses pelatihan model. Langkah-langkah preprocessing yang diterapkan meliputi:

- 1) pengecekan dan penanganan nilai yang hilang (missing values)
- 2) deteksi dan penanganan data duplikat
- 3) pengecekan konsistensi data
- 4) normalisasi format data.

Dari 500 data yang terkumpul, seluruh data dinyatakan lengkap dan siap digunakan untuk proses selanjutnya. Pengecekan missing values dilakukan menggunakan fungsi `isnull().sum()` pada setiap kolom, sementara deteksi duplikat menggunakan fungsi `duplicated()` berdasarkan kombinasi seluruh kolom kecuali Timestamp. Pengecekan konsistensi data mencakup penyeragaman kapitalisasi dan format teks (misalnya variasi penulisan "Ya", "ya", dan "YA"

disatukan menjadi satu kategori) sebelum data diproses lebih lanjut pada tahap Label Encoding.

C. Label Encoding

Algoritma Random Forest memerlukan input dalam bentuk numerik. Oleh karena itu, seluruh variabel kategorikal dikonversi menggunakan teknik Label Encoding. Teknik ini mengubah setiap kategori unik dalam suatu variabel menjadi nilai integer. Sebagai contoh, variabel Jenis Kelamin diubah menjadi nilai 0 untuk Perempuan dan 1 untuk Laki-laki. Demikian pula untuk variabel-variabel lain yang bersifat kategorikal seperti Program Studi, dan variabel bertipe boolean seperti Mengikuti Organisasi Saat Ini, Pernah Mengikuti Organisasi Sebelumnya, Pengaruh Ajakan Teman, dan Memiliki Waktu Luang.

D. Pembagian Data

Dataset yang telah melalui proses preprocessing dan encoding dibagi menjadi dua bagian menggunakan teknik hold-out validation. Sebanyak 80% data (400 data) digunakan sebagai data training untuk melatih model, sedangkan 20% data sisanya (100 data) digunakan sebagai data testing untuk mengevaluasi performa model. Pembagian data dilakukan secara acak (random) dengan stratified sampling untuk memastikan distribusi kelas yang seimbang antara data training dan testing.

E. Algoritma Random Forest

Random Forest tergolong metode ensemble yang menyusun beberapa pohon keputusan secara bersamaan dengan teknik bagging, sehingga setiap pohon dilatih pada sampel data acak yang diambil dengan pengembalian. Tiap pohon hanya memperoleh sebagian data hasil pengembalian acak berulang, dan di setiap percabangan node, pemilihan fitur juga dibatasi pada subset acak tertentu saja. Untuk kasus klasifikasi, keputusan akhir model

diperoleh dari hasil pemugutan suara terbanyak di antara seluruh pohon yang dibangun. Secara matematis, prediksi Random Forest dapat dituliskan sebagai:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_n(x)\}$$

di mana $h_i(x)$ adalah prediksi dari pohon ke- i dan n adalah jumlah total pohon dalam hutan (forest). Variabel utama yang diterapkan dalam penelitian ini adalah $n_estimators = 100$ pohon, $max_depth = \text{None}$ (pohon tumbuh penuh), dan $random_state = 42$ untuk reproducibility.

F. Evaluasi Model

Evaluasi model dilakukan menggunakan empat metrik utama berdasarkan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) dari Confusion Matrix. Rumus masing-masing metrik adalah sebagai berikut:

Accuracy menunjukkan seberapa besar persentase tebakan model yang sesuai dengan kondisi sebenarnya dari seluruh data:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision memperlihatkan ketepatan model saat menyatakan sesuatu sebagai positif:

$$\text{Precision} = TP / (TP + FP)$$

Recall menggambarkan kemampuan model menangkap seluruh kasus positif yang sebenarnya ada:

$$\text{Recall} = TP / (TP + FN)$$

F1-Score adalah perpaduan seimbang antara Precision dan Recall yang dinyatakan dalam satu nilai tunggal:

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

III. HASIL DAN PEMBAHASAN

A. Karakteristik Responden

Dari 500 responden yang berhasil dikumpulkan, distribusi berdasarkan jenis kelamin menunjukkan.

1. Karakteristik Berdasarkan Jenis Kelamin.

Secara keseluruhan, komposisi gender dalam dataset ini cenderung seimbang dengan mayoritas berjenis kelamin laki-laki:

- a. Laki-laki: 265 responden (53%)
 - b. Perempuan: 235 responden (47%)
2. Minat Mengikuti Organisasi.
- Ketika responden ditanya mengenai minat atau keputusan mereka untuk mengikuti organisasi/kegiatan kampus, hasilnya menunjukkan antusiasme yang cukup positif:
- a. Ikut: 271 responden (54.2%)
 - b. Tidak: 229 responden (45.8%)
3. Analisis Jenis Kelamin Terhadap Minat Organisasi.
- Jika kita bedah lebih detail bagaimana ketertarikan berorganisasi ini dibagi berdasarkan gender, polanya terlihat sangat menarik:
1. Kelompok Laki-laki (total 265 orang)
 - a. 136 orang memilih Ikut organisasi
 - b. 129 orang memilih Tidak Ikut organisasi
 2. Kelompok Perempuan (Total 235 orang)
 - a. 135 orang memilih Ikut organisasi
 - b. 100 orang memilih Tidak Ikut organisasi

Mahasiswa perempuan di dalam dataset ini memiliki kecenderungan proporsional yang sedikit lebih tinggi untuk berorganisasi (sekitar 57.4% dari total perempuan memilih Ikut) dibandingkan dengan mahasiswa laki-laki (sekitar 51.3% dari total laki-laki memilih Ikut).

B. Hasil Preprocessing

Proses preprocessing menghasilkan dataset yang bersih dan siap digunakan untuk pelatihan model. Tidak ditemukan nilai yang hilang (missing values) maupun data duplikat pada dataset. Seluruh 8 variabel input berhasil diubah ke dalam format numerik melalui Label

Encoding. Dataset selanjutnya dibagi menjadi 400 data pelatihan dan 100 data pengujian.

1. Pembersihan Nama Kolom.

Pada data mentah (raw data), beberapa kolom memiliki spasi kosong yang tidak konsisten di awal atau di akhir teks (misalnya: ' Program Studi ' atau ' Jenis Kelamin '). Hasilnya semua spasi yang tidak diperlukan berhasil dihapus secara massal, menghasilkan nama variabel yang bersih. Langkah ini penting untuk mencegah terjadinya error KeyError saat pemanggilan variabel pada tahap pemodelan.

2. Seleksi Fitur

Dari seluruh kolom yang tersedia di dataset awal (termasuk Timestamp, Asal Universitas, dll.), dilakukan penyaringan untuk mengambil 5 kolom spesifik yang dinilai memiliki hubungan dengan variabel target. Hasilnya Dataset diperkecil menjadi matriks bersih berukuran 500 baris $\times$ 5 kolom yang fokus pada indikator keaktifan, riwayat organisasi, pengaruh eksternal, ketersediaan waktu, dan keputusan kehadiran.

3. Transformasi Data Kategori

Karena algoritma Machine Learning (Random Forest) berbasis numerik dan tidak dapat memproses data teks secara langsung, dilakukan transformasi menggunakan Label Encoder dari pustaka Scikit-Learn. Data kategorikal biner diubah menjadi format angka (0 dan 1).

4. Pemisahan Variabel Fitur dan Target (X and y Split)

Data yang telah berbentuk numerik penuh kemudian dibagi menjadi dua bagian komponen utama:

- a. Variabel Independen/Fitur (X):
Mengambil 4 kolom pertama yang berisi indikator perilaku responden,

menghasilkan matriks berukuran (500, 4).

- b. Variabel *Dependen/Target* (*y*): Mengambil kolom terakhir (Keputusan Kehadiran) sebagai target prediksi, menghasilkan vektor berukuran (500,).

5. Pembagian Data Evaluasi

Untuk menguji kemampuan generalisasi model secara objektif, 500 data responden tersebut dibagi menggunakan metode Train-Test Split dengan proporsi rasio 80:20 dan dikunci dengan *random_state* = 42:

- a. Data Training (X_{train}, y_{train}): Sebanyak 400 data (80%) dialokasikan untuk fase pembelajaran atau melatih model Random Forest.
- b. Data Testing (X_{test}, y_{test}): Sebanyak 100 data (20%) disisihkan sepenuhnya sebagai instrumen evaluasi (ujian) untuk mengukur performa prediksi model.

C. Hasil Pelatihan Random Forest

Model Random Forest berhasil dilatih menggunakan 400 data training dengan 100 estimator pohon keputusan. Proses pelatihan dilakukan menggunakan framework Scikit-learn pada Python. Setelah proses pelatihan selesai, model diuji coba pada 100 data baru guna melihat sejauh mana ia bisa bekerja dengan data yang belum pernah ditemui sebelumnya.

1. Spesifikasi dan Arsitektur Model

Penelitian ini menggunakan algoritma Random Forest Classifier, sebuah metode ensemble learning berbasis pohon keputusan (decision trees) yang bekerja dengan membangun banyak pohon secara acak untuk meningkatkan akurasi klasifikasi dan mencegah overfitting. Model dilatih menggunakan konfigurasi parameter *n_estimators* = 100 (membangun 100 pohon keputusan) dan dikunci dengan *random_state* = 42 untuk memastikan hasil pengujian dapat direplikasi secara konsisten (reproducible).

2. Pengukuran Tingkat Kepentingan Fitur (Feature Importance)

Setelah model selesai dilatih, dilakukan ekstraksi nilai Feature Importance untuk mengukur kontribusi masing-masing variabel independen dalam menentukan keputusan akhir mahasiswa (Hadir/Tidak Hadir). Hasil pembobotan kontribusi fitur dijabarkan sebagai berikut:

- a. Faktor Utama (Dominan): Ketersediaan waktu luang di luar jam kuliah memiliki bobot tertinggi sebesar 36.0%. Hal ini menunjukkan bahwa kendala waktu atau kesibukan akademik merupakan variabel paling krusial yang dipertimbangkan mahasiswa sebelum menghadiri acara kampus.
- b. Faktor Internal Perilaku: Status keaktifan mengikuti organisasi kampus saat ini memberikan pengaruh sebesar 26.7%, disusul oleh kepemilikan riwayat organisasi sebelumnya sebesar 20.8%. Ini menandakan adanya pengaruh kuat dari habits (kebiasaan) dan pengalaman berorganisasi di masa lalu terhadap tingkat partisipasi mahasiswa saat ini.
- c. Faktor Eksternal (Terendah): Pengaruh ajakan teman memiliki bobot paling kecil, yaitu hanya 16.5%. Temuan ini membuktikan bahwa mahasiswa bertindak secara pragmatis dan mandiri; keputusan untuk hadir tidak didasarkan pada tekanan sosial (peer pressure), melainkan kegunaan acara dan ketersediaan waktu luang mereka sendiri.

3. Ringkasan Evaluasi Performa Klasifikasi (Data Testing)

Setelah melewati fase pelatihan, model diuji menggunakan 100 data testing (20% sisa data yang belum pernah dilihat model) guna mengukur kemampuan generalisasinya. Hasil metrik performa model adalah sebagai berikut:

- a. Akurasi Global (Accuracy): 74.00%
Secara keseluruhan, model mampu menebak dengan benar status

kehadiran 74 dari 100 mahasiswa pada data uji.

- b. Presisi (Precision): 66.67%
Menunjukkan bahwa ketika model memprediksi seorang mahasiswa akan hadir, ketepatan prediksi tersebut mencapai 66,67%, artinya dari seluruh prediksi "Hadir" yang dikeluarkan model, sekitar dua pertiganya sesuai dengan kondisi aktual.
- c. Sensitivitas (Recall): 42.42%
Nilai recall sebesar 42,42% tergolong rendah, yang menunjukkan bahwa model masih kesulitan mengenali sebagian besar mahasiswa yang aktualnya hadir (19 dari 33 kasus Hadir justru terprediksi keliru sebagai Tidak Hadir). Rendahnya nilai ini kemungkinan besar dipengaruhi oleh ketidakseimbangan jumlah kelas pada data pengujian, di mana kelas Tidak Hadir jauh lebih banyak dibandingkan kelas Hadir.
- d. F1-Score: 51.84%
Merupakan nilai keseimbangan rata-rata harmonik antara presisi dan recall pada data uji.

D. Hasil Evaluasi

Untuk menguji keandalan dan kemampuan generalisasi model Random Forest yang telah dilatih, dilakukan pengujian menggunakan data uji (testing data) sebanyak 100 sampel. Berdasarkan perbandingan antara nilai prediksi model dan nilai aktual di lapangan, diperoleh hasil pengukuran metrik evaluasi sebagai berikut:

Tabel 1. Hasil Evaluasi Model Random Forest

Metrik	Nilai
Accuracy	74%
Precision	66%
Recall	42%
F1-Score	51%

Tabel 2. Confusion Matrix Hasil Pengujian Model

	Prediksi: Hadir(1)	Prediksi: Tidak Hadir(0)
Aktual: Hadir (1)	TP = 14	FN = 19
Aktual: Tidak Hadir (0)	FP = 7	TN = 60

Berdasarkan hasil pengujian menggunakan Confusion Matrix, model klasifikasi Random Forest berhasil memprediksi secara akurat sebanyak 14 data yang aktualnya Hadir sebagai Hadir (True Positive), serta 60 data yang aktualnya Tidak Hadir sebagai Tidak Hadir (True Negative). Sebaliknya, terdapat 19 kesalahan prediksi di mana data yang aktualnya Hadir keliru diprediksi sebagai Tidak Hadir (False Negative), dan 7 data yang aktualnya Tidak Hadir salah diprediksi sebagai Hadir (False Positive). Perlu dicatat bahwa dari 100 data uji, hanya 33 data (14 TP + 19 FN) yang aktualnya berkelas Hadir, sedangkan 67 data (7 FP + 60 TN) berkelas Tidak Hadir, sehingga terdapat ketidakseimbangan kelas (class imbalance) pada data pengujian. Kondisi inilah yang menjadi salah satu penyebab nilai recall (42,42%) tergolong rendah, karena model cenderung lebih sering memprediksi kelas mayoritas (Tidak Hadir) sehingga banyak kasus Hadir yang sebenarnya (19 dari 33 data) justru terlewat (FN). Nilai akurasi sebesar 74% pun perlu dimaknai secara hati-hati, sebab pada dataset yang tidak seimbang, akurasi global cenderung bias terhadap kelas mayoritas dan kurang mencerminkan kemampuan model dalam mengenali kelas minoritas. Sejalan dengan

penelitian Jawad dkk. [1] yang juga menerapkan Random Forest pada data pendidikan, penelitian tersebut secara eksplisit menggabungkan Random Forest dengan teknik penyeimbangan data (SMOTE) untuk mengatasi ketidakseimbangan kelas pada dataset pembelajaran daring. Sebagai langkah lanjutan, penerapan SMOTE dapat diuji pada data training saja, bukan pada keseluruhan dataset, untuk menghindari kebocoran informasi ke data testing. Alternatif lain adalah menyesuaikan parameter `class_weight='balanced'` pada Random Forest Classifier. Precision dan recall per kelas sebaiknya dilaporkan secara terpisah pada penelitian selanjutnya, agar dampak penyeimbangan data terhadap kelas minoritas (Hadir) dapat diukur secara langsung. Penting dicatat bahwa akurasi tinggi tidak selalu mencerminkan model yang baik, terutama pada dataset tidak seimbang.

Sebagai perbandingan, Fatunnisa dan Marcos [7] melaporkan akurasi 100% pada prediksi kelulusan siswa SMK, tetapi dataset mereka hanya berisi 160 data dengan distribusi 152 lulus berbanding 8 tidak lulus, dan test set mereka hanya memuat 2 data kelas minoritas. Kondisi ini membuat klaim akurasi sempurna sulit digeneralisasi dan berpotensi mengindikasikan overfitting, bukan performa model yang unggul. Penelitian ini, meskipun mencatat akurasi lebih rendah (74%), menggunakan test set yang lebih besar (100 data, 33 di antaranya kelas minoritas) sehingga evaluasi recall dan precision lebih dapat diandalkan sebagai gambaran kemampuan generalisasi model yang sebenarnya. Dibandingkan dengan Balabied dan Eid [6], yang menerapkan Random Forest untuk deteksi dini underperformance akademik menggunakan dataset OULAD berskala besar (6.519 data uji), penelitian tersebut melaporkan akurasi 90% dengan precision dan recall yang relatif seimbang pada kedua kelas (0,85–0,95). Gap performa ini kemungkinan dipengaruhi oleh dua faktor: ukuran dataset yang jauh lebih besar pada studi tersebut, dan jenis prediktor yang digunakan berupa data perilaku digital (jumlah klik, keterlambatan submisi tugas) yang secara umum

memiliki korelasi lebih kuat dengan variabel target dibanding data kuesioner self-report yang dipakai pada penelitian ini.

Tabel 3. Pemandangan

penelitian	algoritma	akurasi	precision	recall
Penelitian ini	Randoom Forest	74%	66,67%	42,42%
[6] Balabied & Eid	Randoom Forest	90%	85–95%	85–95%
[7] Fatunnisa & Marcos	Randoom Forest	100%	100%	100%

E. Feature Importance

Tabel 4 menampilkan nilai feature importance dari setiap variabel input pada model Random Forest:

Tabel 4. Hasil Analisis Feature Importance

No	Fitur	Nilai Importance
1	Memiliki Waktu Luang	0,36
2	Mengikuti Organisasi Saat Ini	0,26
3	Pernah Mengikuti Organisasi Sebelumnya	0.20
4	Pengaruh Ajakan Teman	0.16

Berdasarkan Tabel 3, ketersediaan waktu luang menjadi faktor paling dominan (36%) dalam menentukan keputusan kehadiran mahasiswa, diikuti oleh status keikutsertaan organisasi saat ini (26,7%) dan riwayat organisasi sebelumnya (20,8%), sedangkan pengaruh ajakan teman memberikan kontribusi terkecil (16,5%). Temuan ini mengindikasikan bahwa keputusan mahasiswa untuk berpartisipasi lebih banyak ditentukan oleh faktor internal (manajemen waktu dan kebiasaan berorganisasi) dibandingkan tekanan sosial dari lingkungan pertemanan, sejalan dengan pembahasan pada bagian Pendahuluan mengenai peran ketersediaan waktu dan pengalaman

berorganisasi [3], [4]. Meskipun demikian, beberapa studi lain menekankan peran signifikan peer influence pada mahasiswa tingkat awal, sehingga karakteristik populasi responden diduga turut memengaruhi pola dominasi faktor yang teridentifikasi. Oleh karena itu, generalisasi hasil ke populasi mahasiswa yang lebih luas perlu dilakukan secara hati-hati.

IV. KESIMPULAN

Studi ini berhasil mengembangkan model prediksi keterlibatan mahasiswa dalam kegiatan kampus dengan memanfaatkan algoritma Random Forest berbasis data kuesioner dari 500 responden mahasiswa. Model yang dibangun dengan konfigurasi 100 pohon keputusan ($n_{\text{estimators}}=100$) menghasilkan akurasi sebesar 74%, dengan nilai precision 66%, recall 42%, dan F1-Score 51% pada data pengujian. Hasil ini menunjukkan bahwa model Random Forest memiliki akurasi sebesar 74% tergolong sedang, terutama karena nilai recall (42%) menunjukkan model belum mampu mengenali sebagian besar mahasiswa yang aktualnya hadir. Nilai akurasi yang tampak baik ini sebagian besar disumbang oleh kemampuan model mengenali kelas mayoritas (Tidak Hadir), bukan performa seimbang pada kedua kelas. Temuan recall rendah ini sejalan dengan Jawad dkk. [1], yang melaporkan tantangan serupa pada data pendidikan dengan distribusi kelas tidak seimbang sebelum menerapkan SMOTE. Meskipun masih terdapat ruang untuk peningkatan, khususnya pada nilai recall dan F1-Score.

Berdasarkan analisis feature importance, ditemukan bahwa ketersediaan waktu luang merupakan faktor paling dominan dengan bobot 36%, diikuti oleh status keikutsertaan dalam organisasi saat ini (26,7%), riwayat organisasi sebelumnya (20,8%), dan pengaruh ajakan teman (16,5%). Temuan ini mengindikasikan bahwa keputusan mahasiswa untuk berpartisipasi dalam kegiatan kampus lebih dipengaruhi oleh faktor internal dan kebiasaan berorganisasi, bukan sekadar tekanan sosial dari lingkungan pertemanan.

Diharapkan hasil dari riset ini bisa dijadikan acuan oleh pihak kampus, terutama Universitas Duta Bangsa Surakarta, saat menyusun langkah-langkah intervensi yang lebih sesuai kebutuhan demi mendorong keikutsertaan mahasiswa secara berkelanjutan. Untuk riset berikutnya, ada baiknya memasukan lebih banyak jenis variable, menambah jumlah responden, dan mencoba teknik oversampling semacam SMOTE supaya performa model, terutama pada kelas minoritas, dapat ditingkatkan.

UCAPAN TERIMA KASIH

Penulis menyampaikan rasa terima kasih yang sedalam-dalamnya kepada Universitas Duta Bangsa Surakarta atas dukungan fasilitas dan sarana yang telah disediakan sepanjang proses penelitian ini berlangsung. Terima kasih juga kami sampaikan kepada Bapak/Ibu dosen pengampu mata kuliah Artificial Intelligence yang telah memberikan bimbingan, petunjuk, dan saran yang sangat berarti dalam penyelesaian penelitian ini. Tidak lupa, kami menyampaikan terima kasih kepada seluruh mahasiswa yang telah bersedia meluangkan waktunya untuk mengisi angket penelitian ini, karena kontribusi tersebut sangat berarti bagi keberhasilan pengumpulan data. Diharapkan penelitian ini mampu memberikan kontribusi bagi kemajuan ilmu pengetahuan, khususnya dalam ranah Educational Data Mining dan penerapan machine learning di lingkungan pendidikan tinggi.

REFERENSI

- [1] K. Jawad, M. A. Shah, and M. Tahir, "Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing," *Sustainability (Switzerland)*, vol. 14, no. 22, Nov. 2022, doi: 10.3390/su142214795.
- [2] P. S. Saputra, "ANALISIS PREDIKTIF DROPOUT MAHASISWA BERDASARKAN KINERJA AKADEMIK SEMESTER AWAL MENGGUNAKAN MACHINE LEARNING," *Jurnal Riset dan Aplikasi Mahasiswa Informatika (JRAMI)*, vol. 07, 2026.
- [3] M. S. W. Putra and H. Baskoro, "Pengaruh keaktifan organisasi, motivasi belajar dan teman sebaya terhadap prestasi belajar mahasiswa program studi manajemen," *Insight Management Journal*, vol. 3, no. 3, pp. 207–216, May 2023, doi: 10.47065/imj.v3i3.264.
- [4] J. Lu *et al.*, "Machine learning analysis of factors affecting college students' academic performance," *Front. Psychol.*, vol. 15, 2024, doi: 10.3389/fpsyg.2024.1447825.

- [5] V. Y. D. Wijaya and G. Brotosaputro, "Penerapan Data Mining Dalam Prediksi Kinerja Akademik Mahasiswa Menggunakan Algoritma Machine Learning."
- [6] S. A. A. Balabied and H. F. Eid, "Utilizing random forest algorithm for early detection of academic underperformance in open learning environments," *PeerJ Comput. Sci.*, vol. 9, 2023, doi: 10.7717/peerj-cs.1708.
- [7] A. Fatunnisa and H. Marcos, "Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 14, no. 1, pp. 101–111, Apr. 2024, doi: 10.34010/jamika.v14i1.12114.
- [8] D. O. E. Wantos and D. S. Bahri, "Analysis and Prediction of Student Academic Performance Using the Random Forest Algorithm," *Jurnal Inotera*, vol. 11, no. 1, pp. 232–239, Apr. 2026, doi: 10.31572/inotera.Vol11.Iss1.2026.ID657.
- [9] R. Hidayat, H. Gultom, and Y. Samudra, "Predicting Student Academic Performance Using Learning Activity Data: A Comparative Study of Random Forest and Decision Tree Models," vol. 4, no. 3, pp. 2829–2529, 2025, doi: 10.52158/rjti.v4i3.4032.
- [10] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. A. Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends Neurosci. Educ.*, vol. 33, Dec. 2023, doi: 10.1016/j.tine.2023.100214.