PERBANDINGAN MODEL *MACHINE LEARNING* UNTUK PREDIKSI *DROP-OUT* MAHASISWA DENGAN ANALISIS *FEATURE IMPORTANCE* BERBASIS STUDI LITERATUR

Athiyyah Nuha Rotifa, Fidela Tertia Alfino,
Puti Chalisa Wardhana, A. Salwa Aurelya Putri, Fathoni
Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Kec. Indralaya,
Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia
fathoni@unsri.ac.id

ABSTRAK

Tingginya angka *drop-out* mahasiswa menjadi isu strategis di perguruan tinggi karena berdampak pada kualitas lulusan, akreditasi, dan reputasi institusi, sehingga diperlukan pendekatan prediktif yang akurat dan interpretatif. Permasalahan utama terletak pada kompleksitas faktor penyebab yang mencakup aspek akademik, sosial, dan psikologis, serta ketidakseimbangan data yang berpotensi menurunkan kinerja model klasifikasi. Penelitian ini bertujuan membandingkan performa beberapa model *machine learning* dan mengidentifikasi faktor dominan dalam prediksi *drop-out* mahasiswa. Metode yang digunakan meliputi pemilihan model melalui studi literatur, pemanfaatan *dataset* publik dari Kaggle, serta penerapan algoritma *Random Forest, Decision Tree,* dan *Logistic Regression* yang dievaluasi menggunakan *accuracy, precision, recall, F1-score,* dan AUC, serta analisis *feature importance* berbasis SHAP. Hasil menunjukkan *Logistic Regression* memiliki performa paling optimal dengan AUC 0,81, *recall* 0,7537, *F1-score* 0,5726, dan *accuracy* 0,7350, yang mencerminkan kemampuan diskriminatif yang lebih stabil dibandingkan model lain. *Random Forest* meskipun menghasilkan *accuracy* tertinggi (0,7805), tidak konsisten pada metrik lainnya, sedangkan *Decision Tree* memiliki performa terendah. Analisis mengungkap bahwa CGPA dan *Stress_Index* merupakan faktor utama, sehingga penelitian ini menegaskan pentingnya keseimbangan antara akurasi dan interpretabilitas dalam pengembangan sistem peringatan dini berbasis data di perguruan tinggi.

Kata kunci : *Drop-out, Random Forest, Decision Tree, Logistic Regression, Feature Importance, SHAP*

1. PENDAHULUAN

Di berbagai negara, tingkat *drop-out* (DO) mahasiswa menjadi masalah penting di perguruan tinggi karena berdampak pada akreditasi dan reputasi lembaga. Misalnya, pada penelitian yang menjelaskan bahwa jumlah mahasiswa yang *drop-out* berpengaruh langsung terhadap akreditasi Universitas XYZ, sehingga diperlukan prediksi lebih awal untuk mencegah penurunan kualitas [1]. Seiring dengan makin banyaknya data akademik, pendekatan ML (*Machine learning*) semakin banyak digunakan untuk merancang sistem peringatan dini (*early warning*) bagi mahasiswa yang berisiko *drop-out* [2]. Penelitian global juga membuktikan bahwa *drop-out* merupakan masalah krusial dalam pendidikan tinggi, sehingga teknologi *machine learning* menjadi solusi potensial untuk membantu institusi mengambil tindakan yang lebih cepat [3].

Prediksi *drop-out* mahasiswa menghadapi berbagai tantangan data. Pertama, penyebab *drop-out* yang sangat kompleks seperti faktor ekonomi, sosial, dan akademik yang bervariasi di masing-masing konteks. Kedua, data pendidikan sering kali tidak seimbang seperti jumlah mahasiswa yang *drop-out* jauh lebih kecil daripada yang lanjut, sehingga perlu teknik khusus seperti SMOTE untuk menangani kesenjangan tersebut [4]. Ketidakseimbangan ini berpotensi membuat model bias ke kelas mayoritas dan menyulitkan deteksi mahasiswa berisiko tinggi. Selain itu, penelitian juga mengidentifikasi fitur dominan yang mempengaruhi *drop-out*, seperti jumlah absen, IPK rendah, atau keterlambatan penyelesaian tugas akhir [1]. Hal-hal tersebut menimbulkan risiko respons terlambat jika model tidak dirancang dengan matang.

Berbagai penelitian sebelumnya telah memanfaatkan pendekatan *machine learning* untuk memprediksi kemungkinan terjadinya *drop-out* mahasiswa dengan menggunakan data akademik maupun non-akademik. Beberapa algoritma yang sering dimanfaatkan dalam penelitian meliputi *Decision Tree, Random Forest, Naïve Bayes,* dan *Support Vector Machine* yang dibandingkan untuk memperoleh model prediksi dengan performa terbaik. Selain itu, beberapa studi juga menerapkan teknik penanganan ketidakseimbangan data serta seleksi fitur untuk meningkatkan kinerja model klasifikasi [5], [1]. Walaupun demikian, mayoritas penelitian cenderung menitikberatkan pada upaya meningkatkan tingkat akurasi maupun membandingkan kinerja antar algoritma, sedangkan kajian yang lebih mendalam mengenai pentingnya masing-masing faktor yang mempengaruhi *drop-out* mahasiswa masih belum banyak dibahas.

Meskipun sejumlah penelitian sebelumnya telah membuktikan kemampuan algoritma *machine*

learning dalam memprediksi kemungkinan terjadinya *drop-out* mahasiswa, akan tetapi penelitian yang secara khusus mengkaji analisis *feature importance* dari berbagai model masih relatif terbatas. Sebagian penelitian lebih berfokus pada evaluasi kinerja model klasifikasi untuk mengidentifikasi *drop-out* mahasiswa seperti penggunaan *Random Forest* maupun *Gradient Boosting* dalam memprediksi status mahasiswa berdasarkan data akademik atau data pendaftaran mahasiswa [2], [3]. Sejalan dengan hal tersebut, diperlukan penelitian yang tidak hanya berfokus pada perbandingan kinerja model *machine learning*, tetapi juga melakukan analisis pada tingkat kepentingan fitur guna menentukan variabel yang paling berpengaruh terhadap risiko *drop-out* mahasiswa.

Berdasarkan keterbatasan tersebut, penelitian ini mengusulkan pendekatan analisis prediksi *drop-out* mahasiswa melalui perbandingan beberapa model *machine learning* serta menerapkan penerapan analisis *feature importance* menggunakan interpretasi SHAP untuk mengidentifikasi faktor-faktor yang paling berpengaruh. Pendekatan ini tidak hanya berfokus pada pencapaian performa prediksi yang optimal, tetapi juga memberikan pemahaman yang lebih mendalam mengenai kontribusi setiap variabel dalam proses prediksi. Dengan demikian, analisis ini diharapkan dapat memberikan pemahaman yang lebih menyeluruh terkait faktor-faktor yang memengaruhi risiko *drop-out* mahasiswa.

Penelitian ini bertujuan membandingkan beberapa model *machine learning* dalam memprediksi kemungkinan *drop-out* mahasiswa serta menganalisis tingkat kepentingan fitur yang mempengaruhi hasil prediksi. Fokus utama penelitian ini adalah mengidentifikasi model *machine learning* yang memiliki performa terbaik sekaligus menentukan faktor-faktor yang paling berpengaruh dalam memprediksi potensi *drop-out* mahasiswa. Hasil penelitian ini diharapkan dapat berperan dalam pengembangan penerapan *machine learning* pada bidang pendidikan, khususnya dalam mendukung upaya deteksi dini mahasiswa yang berpotensi mengalami *drop-out*.

2. TINJAUAN PUSTAKA

Fenomena mahasiswa *drop-out* menjadi isu krusial dalam pendidikan tinggi karena berdampak langsung pada keberhasilan studi mahasiswa serta kinerja institusi, termasuk aspek akreditasi dan reputasi apabila angka *drop-out* cukup tinggi [1]. Berbagai penelitian menunjukkan bahwa keputusan mahasiswa untuk menghentikan studi bukan disebabkan oleh satu faktor tunggal, melainkan merupakan fenomena multidimensional yang melibatkan interaksi berbagai faktor [6]. Faktor akademik sering menjadi indikator utama, seperti rendahnya IPK, pengulangan mata kuliah, kegagalan pengisian KRS, serta tunggakan pembayaran akademik [1]. Selain itu, faktor non-akademik seperti kondisi ekonomi, motivasi belajar, dan lingkungan sosial juga turut berkontribusi terhadap risiko *drop-out* [7]. Beberapa studi juga menegaskan bahwa rendahnya kehadiran dan ketidaksesuaian capaian akademik dengan standar institusi dapat memperbesar kemungkinan mahasiswa tidak melanjutkan studi [8]. Sehingga, pemahaman terhadap fenomena *drop-out* perlu dilakukan secara komprehensif dengan mempertimbangkan keterkaitan berbagai faktor yang memengaruhi keberlanjutan pendidikan mahasiswa.

Penelitian ini menggunakan *dataset* dari Kaggle, *dataset* publik yang banyak dimanfaatkan dalam studi perbandingan kinerja algoritma *machine learning*, baik pada bidang kesehatan seperti klasifikasi penyakit [9], [10], [11]. Selain itu, pada penelitian analisis teks, *dataset* Kaggle juga banyak dimanfaatkan dalam penelitian deteksi ujaran kebencian maupun analisis sentimen pada media sosial dengan berbagai algoritma [12], [13]. Namun, penggunaannya untuk prediksi *drop-out* mahasiswa dan membandingkan beberapa model *machine learning* masih terbatas terutama untuk *dataset* yang digunakan dalam penelitian ini.

Penelitian terdahulu telah membandingkan berbagai algoritma *machine learning* untuk memprediksi *drop-out* mahasiswa guna memperoleh model dengan performa terbaik. Penelitian analisis prediktif *drop-out* mahasiswa menunjukkan bahwa *Random Forest* mampu mencapai akurasi sebesar 70,16% dan menunjukkan kinerja yang lebih baik dibandingkan *Gradient Boosting* dalam memprediksi *drop-out* mahasiswa berdasarkan kinerja akademik semester awal [3]. Hasil serupa juga ditunjukkan dalam penelitian yang membandingkan *Random Forest*, *Gradient Boosting*, dan *Decision Tree* dalam prediksi *drop-out*, dimana *Random Forest* memberikan akurasi tertinggi sebesar 99,67% [1]. Selain itu, *Random Forest* dilaporkan juga memiliki kinerja lebih baik dibandingkan *XGBoost* dengan akurasi 80,56% dalam memprediksi *student attrition* [14].

Pendekatan pohon keputusan seperti *Decision Tree*, C4.5, dan C5.0 banyak digunakan dalam prediksi *drop-out* dengan kinerja klasifikasi yang tinggi, bahkan mencapai akurasi hingga 94,30% dan 98,67% pada beberapa studi [1],[15],[16]. Di sisi lain, model berbasis regresi juga dimanfaatkan dalam sistem peringatan dini *drop-out*, di mana *Logistic Regression* terbukti memiliki performa yang baik pada tahap awal prediksi risiko mahasiswa sebelum tersedia data akademik yang lebih lengkap [2]. Temuan-temuan tersebut menunjukkan bahwa algoritma *Random Forest* dan *Decision Tree* di beberapa penelitian memberikan performa yang kompetitif pada kasus prediksi *drop-out* mahasiswa, sementara *Logistic Regression* tetap relevan sebagai model pembanding dalam pengembangan sistem prediksi berbasis *machine learning*.

Feature importance merupakan teknik dalam *machine learning* yang digunakan untuk menilai kontribusi masing-masing fitur terhadap hasil prediksi

model, sehingga dapat mengidentifikasi faktor yang paling berpengaruh terhadap variabel target. Sejumlah penelitian menunjukkan bahwa *Random Forest* secara konsisten memiliki performa yang unggul dibandingkan metode lain sekaligus mampu mengidentifikasi fitur penting secara efektif. Pada kasus prediksi atrisi karyawan, *Random Forest* mencapai akurasi 93% dan AUC 0,98, dengan fitur seperti *Relationship Satisfaction*, *Work-Life Balance*, dan *Age* sebagai faktor dominan [17]. Hasil serupa juga ditemukan pada klasifikasi tingkat stres pengguna media sosial dan deteksi PCOS, di mana *Random Forest* menunjukkan kinerja yang kompetitif serta mampu mengungkap fitur utama yang memengaruhi hasil prediksi [18], [19].

Konsistensi keunggulan *Random Forest* juga terlihat pada berbagai domain lainnya. Dalam prediksi harga sewa ruko, model ini menghasilkan nilai R^2 sebesar 0,6801 dan MAPE 27,40%, dengan luas bangunan sebagai variabel paling dominan [20]. Pada kasus prediksi pembatalan pemesanan hotel, Random Forest kembali menunjukkan performa terbaik dengan F0.5-Score sebesar 0,8279 dan ROC-AUC 0,9165, di mana *deposit type*, *lead time*, dan *total special request* menjadi faktor utama [21]. Sementara itu, dalam prediksi risiko penyakit kardiovaskular, *Random Forest* mencapai akurasi 99% dan *recall* 100%, mengungguli *Support Vector Machine* dan *Naïve Bayes*, dengan fitur *slope*, *chestpain*, dan *restingBP* sebagai kontributor terbesar [22]. Secara keseluruhan, hasil-hasil tersebut menegaskan bahwa *Random Forest* tidak hanya unggul dari sisi performa prediktif, tetapi juga efektif dalam mengidentifikasi fitur-fitur kunci yang berpengaruh signifikan terhadap berbagai permasalahan klasifikasi.

Kebaruan penelitian ini terletak pada pemilihan dan evaluasi model *machine learning* untuk prediksi *drop-out* mahasiswa berbasis studi literatur. *Random Forest*, *Decision Tree*, dan *Logistic Regression* dipilih karena konsisten menunjukkan kinerja yang baik, meskipun penelitian sebelumnya umumnya hanya berfokus pada akurasi dan belum banyak membahas kontribusi fitur. Masing-masing model memiliki keunggulan, yaitu *Random Forest* pada performa klasifikasi, *Decision Tree* pada kejelasan pola, dan *Logistic Regression* pada nilai *recall*. Oleh karena itu, penelitian ini tidak hanya membandingkan performa ketiganya, tetapi juga menganalisis *feature importance* untuk mengidentifikasi faktor utama yang memengaruhi risiko *drop-out* mahasiswa.

3. METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif dengan memanfaatkan data sekunder sebagai sumber utama. Data dalam penelitian ini berasal dari data "*Student Drop Prediction Dataset" https://www.kaggle.com/datasets/meharshanali/student-drop-out-prediction-dataset* yang didapatkan di situs Kaggle.

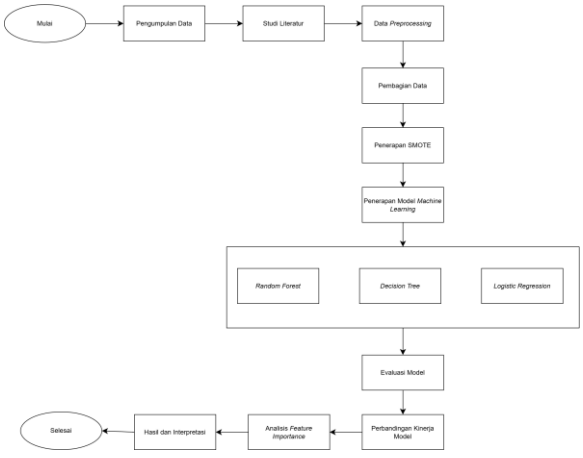
Dataset tersebut terdiri atas 10.000 baris data dengan 19 atribut yang mencakup informasi akademik, demografis, dan sosial mahasiswa, seperti usia, pendapatan keluarga, akses internet, tingkat kehadiran serta nilai akademik seperti GPA dan CGPA. Atribut *drop-out* digunakan sebagai variabel target dalam penelitian ini, sedangkan atribut lainnya berperan sebagai variabel independen.

Jumlah data yang digunakan dalam penelitian ini dinilai memadai untuk penerapan algoritma seperti *Random Forest*, *Decision Tree*, *dan Logistic Regression*. Penelitian terdahulu menyatakan algoritma *Random Forest* memiliki kemampuan dalam menghasilkan prediksi yang konsisten dan akurat, khususnya pada data yang kompleks [23]. Selain itu, *Random Forest* dinilai efektif untuk digunakan pada *dataset* berukuran besar serta mampu mengatasi *overfitting* dan tidak sensitif terhadap data *outlier*, sehingga sesuai untuk diterapkan pada data dengan jumlah besar dan variabel yang beragam [24], [25]. Selain itu, penelitian terdahulu juga menegaskan bahwa algoritma *Decision Tree* efektif untuk menganalisis *dataset* besar. Salah satu penelitian mengungkapkan bahwa pohon keputusan mampu membagi data besar menjadi subset rekam data yang lebih kecil [26]. Sedangkan penelitian lainnya juga mengungkapkan bahwa "*Decision Tree* dapat mengubah data yang sangat besar menjadi sebuah pohon keputusan yang merepresentasikan aturan" [27].

Di sisi lain, *Logistic Regression* juga menunjukkan kinerja yang baik pada *dataset* berukuran besar. Penelitian sebelumnya menyatakan bahwa efisiensi proses pelatihan pada *Logistic Regression* menjadikannya tetap efektif digunakan, khususnya pada *dataset* dengan skala besar [28].

3.1. Alur Penelitian

Penelitian ini dilakukan melalui tahapan yang terstruktur hingga menghasilkan model prediksi *drop-out* mahasiswa berbasis *machine learning*. Rangkaian tahapan tersebut dapat dilihat sebagai berikut.



Gambar 1. Tahapan penelitian

Berdasarkan *flowchart* pada Gambar 1, proses penelitian diawali dengan pengumpulan data dan studi literatur hingga hasil model kemudian dievaluasi dan dianalisis untuk mengidentifikasi faktor yang paling berpengaruh terhadap prediksi *drop-out* mahasiswa. Tahapan penelitian tersebut dijelaskan ke dalam beberapa proses utama sebagai berikut:

a. Data *Prepocessing*
Data *preprocessing* melibatkan serangkaian proses seperti pembersihan data, transformasi, serta penyesuaian format data sehingga data menjadi lebih terstruktur dan mudah diolah [10]. Selain itu, *data preprocessing* ini mencakup deteksi nilai yang hilang, kesalahan data, serta penghapusan atau perbaikan data yang dapat mempengaruhi kualitas analisis. Dengan demikian, data yang dihasilkan menjadi lebih akurat dan siap digunakan dalam proses pemodelan [9]. *Dataset* yang digunakan dalam penelitian ini terdiri dari 10.000 data mahasiswa dengan berbagai atribut akademik dan non-akademik. Pada tahap awal, dilakukan pembersihan data (*data cleaning*) untuk memastikan tidak terdapat inkonsistensi maupun kesalahan format data. Selanjutnya dilakukan penanganan *missing value* pada beberapa atribut seperti *Family_Income*, *Study_Hours_per_Day*, *Stress_Index*, dan *Parental_Education* dengan menggunakan metode tertentu agar tidak mempengaruhi kinerja model. Tahap berikutnya adalah melakukan transformasi data kategorikal menjadi data numerik melalui teknik *encoding*, terutama pada atribut seperti *Gender*, *Internet_Access*, *Part_Time_Job*, dan *Scholarship*, sehingga dapat diproses oleh algoritma *machine learning*. Selain itu, dilakukan pula penyesuaian tipe data pada atribut tertentu, seperti *Attendance_Rate*, agar sesuai dengan format numerik. Selanjutnya, dilakukan penanganan ketidakseimbangan data (*imbalanced data*) menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) [1]. Teknik ini digunakan untuk meningkatkan jumlah data pada kelas minoritas, yaitu mahasiswa yang mengalami *drop-out*, sehingga distribusi data menjadi lebih seimbang.

b. Pembagian Data
Setelah melalui tahap *preprocessing*, *dataset* selanjutnya dibagi menjadi dua bagian, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Pembagian ini bertujuan untuk melatih model serta mengevaluasi kinerjanya dalam memprediksi *drop-out* mahasiswa berbasis *machine learning*.

Tabel 1. Data pelatihan dan pengujian

Jenis Data	Persentase	Jumlah Data
Data Pelatihan	80%	8.000
Data Pengujian	20%	2.000
Total	100%	10.000

Berdasarkan Tabel 1, dapat dilihat bahwa sebagian besar data digunakan sebagai data pelatihan, yaitu sebesar 80% dari total *dataset*, sedangkan 20% sisanya digunakan sebagai data pengujian. Pembagian ini dilakukan untuk memastikan bahwa model memiliki kemampuan generalisasi yang baik dan tidak hanya menghafal data pelatihan [10]. Selanjutnya, untuk mengatasi ketidakseimbangan data, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan hanya pada data pelatihan. Data pengujian tetap menggunakan data asli agar proses evaluasi model dapat mencerminkan performa yang lebih objektif dan menghindari terjadinya kontaminasi data [29].

c. Penerapan Model *Machine learning*
Pada penelitian ini digunakan tiga algoritma *machine learning*, yaitu *Random Forest*, *Decision Tree*, dan *Logistic Regression* untuk melakukan klasifikasi terhadap status *drop-out* mahasiswa. Implementasi model dilakukan menggunakan bahasa pemrograman *Python* dengan dukungan *library Scikit-learn* sebagai *framework* utama dalam pengembangan model, sedangkan seluruh proses eksperimen dijalankan pada lingkungan komputasi berbasis *cloud* menggunakan Google Colab.

- *Random Forest*
Random Forest merupakan algoritma *ensemble learning* yang menggabungkan banyak *decision tree* melalui *majority voting* untuk meningkatkan akurasi dan stabilitas serta mengurangi *overfitting* [29], [30]. Selain itu, *Random Forest* banyak dimanfaatkan baik pada tugas klasifikasi maupun regresi, terutama ketika berhadapan dengan *dataset* berukuran besar dan memiliki banyak variabel [31]. Dalam pembentukannya, *Random Forest* menerapkan *bootstrap sampling* pada data dan pemilihan subset fitur secara acak (*feature bagging*), sehingga menghasilkan variasi antar pohon, menurunkan korelasi, dan lebih efektif dalam menangani data kompleks maupun berdimensi tinggi pada tugas klasifikasi dan regresi [29], [30], [31].
- *Decision Tree*
Decision Tree klasifikasi berbentuk pohon, di mana *node* mewakili atribut, cabang menunjukkan hasil pengujian, dan *leaf node* sebagai keputusan akhir [13]. Model dibangun secara *top-down* dan rekursif dengan memilih atribut terbaik sebagai pemisah data hingga terbentuk kelompok yang semakin homogen [32]. Metode ini unggul karena mampu menghasilkan aturan yang mudah dipahami serta fleksibel dalam menangani atribut kategorikal [13]. Pada algoritma C4.5, pemilihan atribut terbaik ditentukan menggunakan *Information Gain*, yaitu ukuran kemampuan atribut dalam mengurangi ketidakpastian (*entropy*). Atribut dengan nilai

gain tertinggi dipilih sebagai *node* pada setiap tahap pembentukan pohon [32]. Perhitungan *Information Gain* dirumuskan sebagai berikut:

$Gain\ (\mathrm{S,A}) = Entropy\ (s) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i)$ (1)

dengan:

{S1, S2, S3, ..., Si, ..., Sn} = partisi S sejumlah nilai atribut A

N = jumlah atribut A

|*Si*| = Jumlah kasus dalam partisi Si

|*S*| = Total kasus S

Sementara untuk memperoleh nilai entropi sebagai berikut

$Entropy\ (S) = \sum_{i=1}^{n} -p_i * log_2 p_i$ (2)

dengan:

S = jumlah kasus

n = jumlah kasus dalam partisi S

pi = proporsi Si terhadap S

- *Logistic Regression*

  Logistic Regression merupakan metode klasifikasi yang digunakan untuk memodelkan probabilitas keluaran kategorikal, khususnya pada kasus biner, melalui fungsi logistik (*sigmoid*) yang menghasilkan nilai antara 0 hingga 1. Model ini dipilih sebagai *baseline* karena sederhana, efisien secara komputasi, serta mampu memberikan interpretasi yang jelas terhadap hubungan antara variabel *input* dan probabilitas hasil prediksi [21], [33]. Selain itu, model ini mengasumsikan hubungan linier antara variabel *input* dengan *log-odds*, sehingga proses pelatihan relatif cepat dan mudah diterapkan [34]. Secara matematis, *Logistic Regression* memodelkan probabilitas kelas sebagai transformasi dari kombinasi linier variabel *input* menggunakan fungsi *sigmoid*, yang dirumuskan sebagai berikut:

  $p(G = k|X = x) = \frac{1}{1+(exp-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k))}$ (3)

  Model ini memiliki keterbatasan dalam menangkap hubungan non-linier yang kompleks, sehingga dalam banyak penelitian digunakan sebagai model pembanding awal sebelum menerapkan algoritma yang lebih kompleks seperti metode berbasis *ensemble* [21], [34].

d. Validasi Hasil Penelitian

Evaluasi dalam penelitian ini dilakukan melalui validasi performa model untuk mengukur dan membandingkan kinerja beberapa algoritma *machine learning* dalam memprediksi kemungkinan *drop-out* mahasiswa. Penilaian dilakukan menggunakan metrik klasifikasi seperti *accuracy, precision, recall,* dan *F1-score*, yang dilengkapi dengan analisis ROC *curve* serta nilai AUC guna memberikan gambaran kemampuan model secara menyeluruh. Pendekatan ini umum digunakan dalam studi klasifikasi berbasis algoritma seperti *Random Forest, Decision Tree,* dan *Logistic Regression* karena memungkinkan perbandingan kinerja model secara objektif dan terukur. Berikut perhitungan metrik evaluasi tersebut dinyatakan dalam bentuk persamaan [22]..

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ (4)

Accuracy (akurasi) menunjukkan tingkat ketepatan model dengan membandingkan jumlah prediksi yang benar terhadap seluruh data, mencakup kedua kelas sehingga mencerminkan kinerja secara keseluruhan.

$Precision = \frac{TP}{TP+FP}$ (5)

Precision (presisi) menggambarkan tingkat ketepatan model dalam memprediksi kelas positif, yaitu proporsi data yang diprediksi positif dan benar-benar termasuk dalam kelas tersebut.

$Recall = \frac{TP}{TP+FN}$ (6)

Recall (sensitivitas) menunjukkan kemampuan model dalam mengenali seluruh data yang benar-benar termasuk kelas positif.

$F1 - Score = 2\,x\frac{Precision\ x\ Recall}{Precision\ x\ Recall}$ (7)

F1-score merupakan gabungan *precision* dan *recall* dalam bentuk rata-rata harmonis yang mencerminkan keseimbangan keduanya dalam menilai kinerja model.

ROC-AUC digunakan sebagai metrik tambahan karena tidak bergantung pada ambang batas dan mampu menilai kemampuan model dalam membedakan kelas melalui hubungan TPR dan FPR (FPR = FP/(FP+TN)) [2]. Penggunaannya bersama *accuracy, precision, recall,* dan *F1-score* memberikan evaluasi yang lebih menyeluruh, terutama pada data tidak seimbang, sehingga perbandingan performa model lebih objektif [35]. Selanjutnya, dilakukan perbandingan antar model untuk menentukan model terbaik yang akan digunakan dalam analisis *feature importance* guna mengidentifikasi faktor paling berpengaruh terhadap prediksi *drop-out* mahasiswa.

e. Analisis *Feature Importance*

Setelah model terbaik diperoleh, penelitian ini melanjutkan tahap analisis *feature importance* menggunakan pendekatan XAI menggunakan metode SHAP untuk menginterpretasikan hasil prediksi model. Metode ini menghitung kontribusi setiap fitur terhadap *output* pada tiap observasi, sehingga memberikan penjelasan baik secara global maupun lokal, di mana nilai positif menunjukkan peningkatan kecenderungan *drop-out* dan nilai negatif menunjukkan sebaliknya. Secara matematis, nilai SHAP didefinisikan melalui persamaan berikut [36]:

$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$ (8)

dengan :

ϕ_i = nilai SHAP fitur ke-i,

S = subset fitur yang tidak mengandung fitur ke-i,

$f(S)$ = nilai prediksi model berdasarkan subset fitur S.

Pendekatan SHAP digunakan untuk mengidentifikasi fitur dominan melalui analisis global dan lokal, dengan rata-rata nilai absolut SHAP untuk menentukan tingkat kepentingan fitur serta kontribusi tiap fitur pada setiap prediksi. Hasilnya divisualisasikan melalui *feature importance* dan SHAP *summary plot*, sehingga meningkatkan transparansi, interpretabilitas, dan pemahaman hubungan antara fitur dan hasil prediksi [37].

4. HASIL DAN PEMBAHASAN

4.1. Hasil Data *Prepocessing*

Pada tahap data *preprocessing*, dilakukan penanganan *missing value*, transformasi data kategorikal, pemisahan fitur dan target, serta normalisasi data. Tahapan ini disusun untuk menjaga konsistensi data serta mencegah terjadinya data *leakage* antara data pelatihan dan data pengujian [38].

Nilai yang hilang pada atribut *Family_Income*, *Study_Hours_per_Day*, *Stress_Index*, dan *Parental_Education* diisi menggunakan nilai rata-rata (*mean*) dari masing-masing atribut, sedangkan atribut kategorikal diisi menggunakan nilai yang paling sering muncul (modus). Selanjutnya, data kategorikal seperti *Gender*, *Internet_Access*, *Part_Time_Job*, dan *Scholarship* dikonversi menjadi data numerik menggunakan teknik Label *Encoding*. Tahap berikutnya adalah pemisahan antara fitur (*feature*) dan target (label), di mana seluruh atribut selain *drop-out* digunakan sebagai variabel independen (X), sedangkan atribut *drop-out* dijadikan sebagai variabel dependen (y).

Dilakukan normalisasi data pada seluruh fitur numerik menggunakan metode *StandardScaler* yang bertujuan untuk menyamakan skala antar fitur sehingga memiliki distribusi dengan rata-rata mendekati nol dan standar deviasi mendekati satu.

4.2. Pembagian Data dan SMOTE

Pada tahap ini, *dataset* yang telah melalui proses *preprocessing* selanjutnya dibagi menjadi dua bagian, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Pembagian ini bertujuan untuk melatih model serta mengevaluasi kinerjanya dalam memprediksi potensi *drop-out* mahasiswa. Proses pembagian data dilakukan dengan rasio 80:20 menggunakan metode *stratified sampling*.

Selanjutnya, dilakukan analisis distribusi kelas pada data pelatihan sebelum penerapan teknik *penyeimbangan* data. Hal ini bertujuan untuk mengetahui apakah terdapat ketidakseimbangan antara jumlah kelas mayoritas dan minoritas [1].

Tabel 2. Distribusi kelas sebelum SMOTE

Kelas	Jumlah
Tidak *Drop-out*	6117
Drop-out	1883

Berdasarkan Tabel 2, terlihat bahwa jumlah data pada kelas tidak *drop-out* jauh lebih besar dibandingkan dengan kelas *drop-out*. Kondisi ini menunjukkan adanya ketidakseimbangan data (*imbalanced data*), yang berpotensi menyebabkan model cenderung bias terhadap kelas mayoritas dan kurang mampu mengenali pola pada kelas minoritas.

Untuk mengatasi permasalahan tersebut, diterapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) pada data pelatihan. Teknik ini bekerja dengan cara menambahkan data sintetis pada kelas minoritas hingga jumlahnya seimbang dengan kelas mayoritas [29].

Tabel 3. Distribusi kelas setelah SMOTE

Kelas	Jumlah
Tidak *Drop-out*	6117
Drop-out	6117

Berdasarkan Tabel 3, dapat dilihat bahwa jumlah data pada kedua kelas menjadi seimbang untuk kelas tidak *drop-out* dan *drop-out*. Selain itu, penerapan SMOTE hanya dilakukan pada data pelatihan, sedangkan data pengujian tetap *menggunakan* data asli. Hal ini bertujuan untuk menjaga objektivitas dalam proses evaluasi model serta menghindari terjadinya kontaminasi data (data *leakage*) [38].

4.3. Hasil Evaluasi

Evaluasi kinerja mode pada penelitian ini dilakukan untuk mengetahui seberapa baik model mampu mengklasifikasikan data mahasiswa yang berpotensi mengalami *drop-out* menggunakan metrik *accuracy*, *precesion*, *recall* dan *FI-score*, sebagaimana disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi Setiap Model

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.7805	0,5341	0,5308	0,5325
Decision Tree	0.6885	0,3721	0,4692	0,4150
Logistic Regression	0.7350	0,4616	0,7537	0,5726

Berdasarkan hasil evaluasi, model *Random Forest* memperoleh nilai *accuracy* sebesar 0,7805 dengan *precision* 0,5341, *recall* 0,5308 serta *F1-score* sebesar 0,5325. Hasil ini menyatakan algoritma *Random Forest* menghasilkan kinerja yang cukup baik dalam proses klasifikasi data secara keseluruhan dengan tingkat ketepatan yang relatif tinggi.

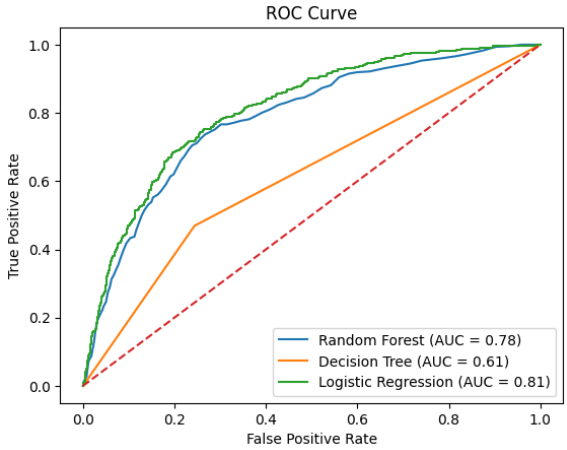
Pada model *Decision Tree* diperoleh nilai *accuracy* sebesar 0,6885, *precision* 0,3721, *recall* 0,4692, dan *F1-score* sebesar 0,4150. Dibandingkan dengan model lainnya, *Decision Tree* menunjukkan kinerja yang lebih rendah, terutama pada nilai *precision* dan *F1-score*, yang mencerminkan keterbatasan mengklasifikasikan kelas positif secara akurat.

Adapun model *Logistic Regression* yang memperoleh nilai *accuracy* sebesar 0,7350, *precision* sebesar 0,4616, *recall* sebesar 0,7537, dan *F1-score* sebesar 0,5726. Nilai *recall* yang cukup tinggi

menunjukkan bahwa *Logistic Regression* memiliki kemampuan yang baik dalam mengidentifikasi mahasiswa yang berpotensi mengalami *drop-out*.

4.4. ROC *Curve* & AUC

Analisis ROC *Curve* dan nilai AUC digunakan karena tidak bergantung pada satu nilai ambang tertentu, sehingga mampu menggambarkan performa model yang lebih komprehensif, seperti yang ditunjukkan pada Gambar 2.



Gambar 2. Kurva ROC dan Nilai AUC Setiap Model

Berdasarkan kurva yang dihasilkan, model *Logistic Regression* menunjukkan nilai AUC tertinggi yaitu 0,81, sehingga menunjukkan kemampuan mengklasifikasi sangat baik dalam membedakan mahasiswa yang berpotensi mengalami *drop-out* dan yang tidak. Sementara itu model *Random Forest* memperoleh nilai AUC sebesar 0,78, yang juga menunjukkan performa cukup baik. Di sisi lain, *Decision Tree* hanya menghasilkan nilai AUC sebesar 0,61, yang menandakan bahwa kemampuannya dalam membedakan kedua kelas masih tergolong rendah dibandingkan model lainnya.

Hasil ini menunjukkan bahwa *Logistic Regression* memiliki kemampuan diskriminatif paling unggul di antara ketiga model yang diuji, sejalan dengan nilai *recall* dan *F1-score* yang relatif lebih tinggi. Namun *Random Forest* juga tetap memperlihatkan performa yang kompetitif dengan nilai AUC yang cukup tinggi, khususnya dalam hal akurasi keseluruhan. Sebaliknya, *Decision Tree* menunjukkan kinerja yang kurang optimal, baik dari sisi metrik evaluasi maupun kemampuan diskriminatif berdasarkan kurva ROC.

4.5. Perbandingan Model

Perbandingan kinerja model dilakukan untuk mengidentifikasi algoritma yang memiliki performa paling optimal dalam memprediksi *drop-out* mahasiswa dengan mempertimbangkan nilai *accuracy* dan *F1-score* sebagai indikator utama.

Tabel 5. Hasil Perbandingan Model *Machine learning*

Model	Accuracy	F1-Score
Random Forest	0.7805	0,5325
Decision Tree	0,6885	0,4150
Logistic Regression	0.7350	0,5726

Berdasarkan hasil pembandingan yang disajikan pada Tabel 5, model *Random Forest* mengindikasikan kemampuan yang baik dalam melakukan klasifikasi secara umum dengan *accuracy* tertinggi sebesar 0,7805. Namun dengan *F1-score* sebesar 0,5325, menggambarkan bahwa keseimbangan antara *precision* dan *recall* pada model ini masih belum optimal. Sementara itu, model *Decision Tree* menunjukkan performa yang lebih rendah dengan nilai *accuracy* sebesar 0,6885 dan *F1-score* sebesar 0,4150, sehingga kemampuan model ini dalam mengklasifikasikan data serta mendeteksi kelas positif masih terbatas. Di sisi lain, *Logistic Regression* memperoleh nilai *accuracy* sebesar 0,7350 dengan *F1-score* tertinggi, yaitu 0,5726, sehingga menunjukkan keseimbangan yang lebih baik antara *precision* dan *recall* dalam mendeteksi mahasiswa berpotensi *drop-out*, yang menjadi aspek lebih penting dibandingkan sekadar akurasi keseluruhan.

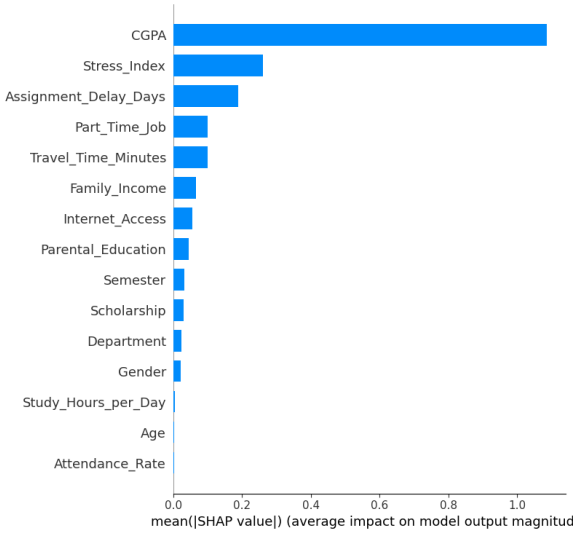
Hasil tersebut menunjukkan bahwa *Logistic Regression* merupakan model terbaik dalam penelitian ini dengan *F1-score* sebagai dasar penentuan. Oleh karena itu, model *Logistic Regression* digunakan pada tahap analisis *feature importance* untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap prediksi *drop-out* mahasiswa.

4.6. Hasil *Feature Importance*

Berdasarkan hasil evaluasi, *Logistic Regression* menunjukkan performa paling optimal dibandingkan *Random Forest* dan *Decision Tree*, karena mampu menjaga keseimbangan antara *precision* dan *recall*, khususnya pada data tidak seimbang. *Logistic Regression* lebih unggul dari sisi kestabilan klasifikasi serta kemampuan diskriminatif yang ditunjukkan oleh nilai AUC. Temuan ini konsisten dengan penelitian sebelumnya yang menyatakan bahwa *Logistic Regression* efektif untuk klasifikasi dan pengembangan sistem peringatan dini karena kestabilan performa dan keseimbangan metrik evaluasinya [2], [39]. Oleh karena itu, model ini dapat dianggap sebagai pendekatan yang relevan dan andal dalam pemodelan klasifikasi pada data pendidikan yang bersifat terstruktur.

Setelah model terbaik diperoleh, analisis *feature importance* dilakukan menggunakan *Logistic Regression* dengan pendekatan XAI melalui metode SHAP untuk menilai kontribusi setiap variabel terhadap prediksi *drop-out*, termasuk penilaian kontribusi global setiap fitur sebagaimana ditunjukkan pada Gambar 3. Terlihat bahwa variabel CGPA memiliki nilai rata-rata absolut SHAP tertinggi, sehingga menjadi faktor paling dominan dalam model dan berkontribusi terhadap prediksi *drop-out*. Hal ini

menegaskan peran utama performa akademik dalam keberlangsungan studi mahasiswa.



Gambar 3. *Feature Importance*

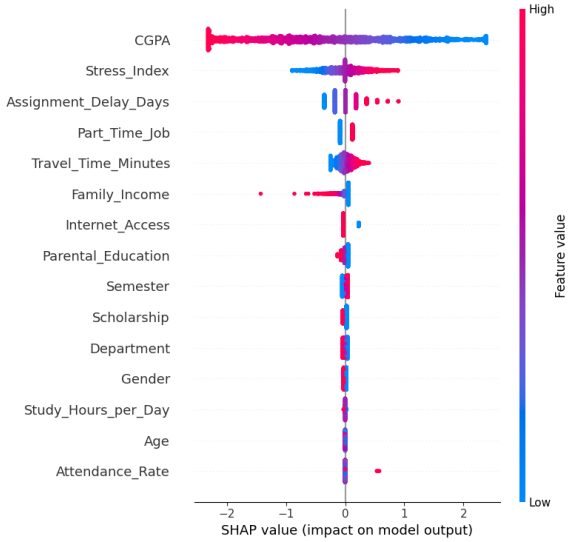
Temuan ini sejalan dengan penelitian yang menunjukkan bahwa indeks prestasi mahasiswa dari semester awal hingga pertengahan studi memiliki peran penting dalam memprediksi keberhasilan kelulusan karena mencerminkan konsistensi capaian akademik [40]. Selain itu, penelitian juga menemukan bahwa GPA memiliki hubungan yang signifikan secara statistik terhadap risiko *drop-out* ($p < 0{,}001$), di mana mahasiswa dengan GPA rendah cenderung memiliki probabilitas *drop-out* yang lebih tinggi [41]. Dengan demikian, GPA tidak hanya berfungsi sebagai indikator kinerja akademik, tetapi juga dapat digunakan sebagai indikator awal untuk mengidentifikasi potensi risiko *drop-out*.

Selanjutnya, variabel *Stress_Index* menempati posisi kedua sebagai faktor yang berpengaruh terhadap prediksi *drop-out*. Hal ini menunjukkan bahwa tingkat stres berkontribusi signifikan terhadap peningkatan risiko *drop-out* karena dapat menurunkan fokus, motivasi, dan kemampuan akademik. Temuan ini sejalan dengan penelitian yang menyatakan bahwa stres dan *burnout* berdampak pada penurunan performa belajar serta meningkatnya kecenderungan mahasiswa untuk tidak melanjutkan studi [42]. Aspek psikologis menjadi faktor penting yang tidak dapat diabaikan dalam analisis risiko *drop-out.*

Selain itu, variabel *Assignment_Delay_Days*, *Part_Time_Job*, dan *Travel_Time_Minutes* berkontribusi cukup signifikan, yang mencerminkan kendala manajemen waktu, beban aktivitas non-akademik, serta hambatan kehadiran dalam perkuliahan. Hal ini menunjukkan bahwa faktor operasional dan lingkungan turut berperan dalam memengaruhi keberlangsungan studi mahasiswa. Sementara itu, variabel seperti *Family_Income* memberikan kontribusi pada tingkat moderat. Hal ini mengindikasikan bahwa faktor ekonomi dan beban aktivitas tambahan tetap memiliki pengaruh terhadap risiko *drop-out*, meskipun tidak sebesar variabel utama. Sedangkan *Attendance_Rate*, *Age*, dan *Study_Hours_per_Day* menunjukkan kontribusi yang relatif rendah, sehingga pengaruhnya terhadap prediksi model cenderung terbatas dalam konteks data yang digunakan.

Secara keseluruhan, hasil analisis *feature importance* menunjukkan bahwa prediksi *drop-out* mahasiswa dipengaruhi oleh kombinasi faktor akademik, psikologis, serta kondisi eksternal. Dominasi variabel CGPA dan *Stress_Index* menegaskan bahwa keberhasilan akademik dan kondisi mental merupakan dua aspek utama dalam mengidentifikasi mahasiswa berisiko tinggi *drop-out.*

Analisis *SHAP summary plot* juga dilakukan untuk mengevaluasi kontribusi keseluruhan setiap fitur terhadap prediksi model secara komprehensif Visualisasi ini menampilkan distribusi nilai SHAP, di mana setiap titik mewakili satu data dan warna menunjukkan nilai fitur, sehingga arah serta kekuatan pengaruhnya terhadap peningkatan atau penurunan risiko *drop-out* dapat diinterpretasikan dengan jelas.



Gambar 4. *SHAP Summary Plot*

Berdasarkan SHAP *summary plot* pada Gambar 4, terlihat bahwa variabel CGPA memiliki pengaruh paling dominan dalam membentuk prediksi model. Temuan ini sejalan dengan penelitian yang mengelompokkan mahasiswa berpotensi *drop-out* dan mengidentifikasi IPK sebagai faktor utama dalam mengelompokkan mahasiswa berpotensi *drop-out*, bersama variabel akademik lainnya seperti jumlah SKS dan status keaktifan [43]. Hasil serupa juga dikemukakan oleh penelitian yang menunjukkan bahwa mahasiswa dengan IPK rendah cenderung berada pada kelompok berisiko tinggi *drop-out*, khususnya ketika disertai ketidakterpenuhinya capaian akademik pada mata kuliah tertentu [44].

Variabel *Stress_Index* menempati posisi berikutnya dengan pola distribusi yang menunjukkan bahwa nilai stres tinggi (warna merah) cenderung

berada pada sisi SHAP positif, menegaskan bahwa tekanan akademik dapat memicu kecenderungan meningkatnya prediksi *drop-out*. Temuan ini sejalan dengan studi yang menyatakan bahwa kemampuan mengelola stres berkaitan dengan *academic resilience*, di mana ketidakmampuan menghadapi tekanan dapat menurunkan motivasi dan performa belajar serta meningkatkan kemungkinan menghentikan studi [45].

Selain itu, variabel *Assignment_Delay_Days* menunjukkan kecenderungan bahwa nilai yang lebih tinggi berkorelasi dengan nilai SHAP positif, yang mengindikasikan bahwa keterlambatan dalam penyelesaian tugas berhubungan dengan peningkatan risiko *drop-out*. Variabel *Part_Time_Job* dan *Travel_Time_Minutes* juga berkontribusi positif, menandakan bahwa aktivitas non-akademik dan hambatan mobilitas dapat memengaruhi keberlangsungan studi. Temuan ini sejalan dengan penelitian yang menyatakan bahwa keterlibatan dalam pekerjaan paruh waktu dapat mengurangi fokus belajar dan meningkatkan risiko ketidaktercapaian target akademik [46].

Sementara itu, variabel seperti *Family_Income*, *Internet_Access*, dan *Parental_Education* menunjukkan kontribusi moderat. Berdasarkan konteks data yang digunakan, faktor ekonomi dan akses pendukung tidak menjadi penentu utama dalam prediksi *drop-out*. Secara keseluruhan, hasil SHAP *summary plot* menegaskan bahwa interaksi antara faktor akademik dan psikologis memiliki peran yang lebih dominan dalam membentuk risiko *drop-out* dibandingkan faktor eksternal lainnya.

Hasil analisis secara keseluruhan menunjukkan bahwa faktor akademik, khususnya CGPA, merupakan determinan utama dalam memprediksi risiko *drop-out*, diikuti oleh faktor psikologis berupa *Stress_Index*. Temuan ini konsisten dengan sebagian besar penelitian sebelumnya yang menempatkan IPK sebagai indikator kunci keberhasilan studi. Namun demikian, penelitian ini memberikan kontribusi tambahan dengan menunjukkan bahwa variabel non-akademik seperti stres dan keterlambatan tugas memiliki peran yang relatif kuat, sehingga memperluas perspektif bahwa risiko *drop-out* tidak hanya ditentukan oleh capaian akademik, tetapi juga oleh kondisi psikologis dan perilaku belajar mahasiswa. Dengan demikian, hasil ini tidak hanya mengonfirmasi temuan terdahulu, tetapi juga memperkuat urgensi pendekatan multidimensional dalam sistem peringatan dini.

Dari sisi implikasi, hasil ini dapat digunakan sebagai dasar pengembangan kebijakan institusi, khususnya dalam merancang intervensi berbasis data seperti monitoring performa akademik secara berkala, penyediaan layanan konseling untuk manajemen stres, dan pendampingan mahasiswa dengan kendala akademik agar indikator-indikator tersebut dapat menjadi acuan dalam antisipasi terjadinya *drop-out*. Penelitian ini telah mencapai tujuan dengan membandingkan tiga model *machine learning* dan mengidentifikasi faktor utama risiko *drop-out*, namun masih terbatas pada penggunaan data sekunder dan belum mempertimbangkan aspek longitudinal.

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa perbandingan tiga algoritma *machine learning*, yaitu *Random Forest*, *Decision Tree*, dan *Logistic Regression*, mampu memberikan gambaran kinerja model dalam memprediksi risiko *drop-out* mahasiswa secara komprehensif. Berdasarkan hasil evaluasi, *Logistic Regression* menjadi model paling optimal karena memiliki keseimbangan terbaik pada metrik evaluasi, khususnya *F1-score* dan AUC, meskipun *Random Forest* mencatat akurasi tertinggi. Analisis *feature importance* berbasis SHAP mengidentifikasi bahwa CGPA dan *Stress_Index* merupakan faktor paling dominan, disertai kontribusi variabel lain seperti *Assignment_Delay_Days*, *Part_Time_Job* dan *Travel_Time_Minutes*, sehingga menegaskan bahwa risiko *drop-out* dipengaruhi oleh kombinasi faktor akademik, psikologis, dan lingkungan. Temuan ini mendukung pengembangan sistem peringatan dini untuk mitigasi terjadinya *drop-out* yang tidak hanya akurat tetapi juga interpretatif. Untuk penelitian selanjutnya, disarankan penggunaan data yang lebih kontekstual serta pendekatan longitudinal agar hasil prediksi menjadi lebih representatif dan aplikatif.

DAFTAR PUSTAKA

[1] T. A. Marzuqi, E. Kristiani, and Marcel, "Prediksi Mahasiswa Drop-Out Di Universitas XYZ," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, pp. 1345–1350, 2024, doi: 10.25126/jtiik.2024118689.

[2] F. A. Putra, S. Mirajdandi, Nandra, B. Okmarizal, and S. Mulyanda, "Prediksi Dropout Mahasiswa : Early-Warning Berbasis Enrollment dengan *Machine learning*," *J. Fasilkom Univ. Muhamma*, vol. 15, no. 3, pp. 465–473, 2025.

[3] P. S. Saputra, "Analisis Prediktif Dropout Mahasiswa Berdasarkan Kinerja Akademik Semester Awal Menggunakan *Machine learning*," *J. Ris. dan Apl. Mhs. Inform.*, vol. 07, no. 01, pp. 164–171, 2026.

[4] F. L. Hanis, U. Khaira, and D. Arsa, "Klasifikasi Status Mahasiswa Berisiko Drop OutMenggunakan Decision Tree C5.0 dengan Seleksi Fitur," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. October, pp. 1318–1330, 2025.

[5] M. T. Anwar, D. Rianditha, and A. Permana, "Perbandingan Performa Model Data Mining untuk Prediksi Dropout Mahasiswa," vol. 2, pp. 87–94, 2021, doi: 10.52330/jtm.v19i2.34.

[6] H. P. Sukmarasa, E. Maulana, M. R. Ardhiansyah, O. Hijuzaman, and S. Info, "Model Multivariat untuk Memprediksi Angka Putus Sekolah Mahasiswa di Perguruan Tinggi di Wilayah Purwakarta," *J. Manaj. Pendidik.*, vol. 11, no. 1, pp. 965–974, 2026.

[7] A. S. Wahyuni, J. La Fua, and Rasmi, "Analisis Faktor Penyebab Putus Sekolah pada Siswa Sekolah Dasar: Studi Kasus di Desa Mola Bahari, Indonesia," *DINIYAH J. Pendidik. Dasar*, vol. 5, no. 2, pp. 124–134, 2024.

[8] R. Hidayat, M. Haris, and Z. K. Simbolon, "Implementasi Metode Support Vector Machine (SVM) Pada Klasifikasi Drop Out (DO) Mahasiswa," *J. Infomedia Tek. Inform. Multimedia, dan Jar.*, vol. 9, pp. 51–59, 2024.

[9] K. Akmal, A. Faqih, and F. Dikananda, "Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 470–477, 2023, doi: 10.36040/jati.v7i1.6367.

[10] R. Harahap, M. Irpan, M. A. Dinata, L. Efrizoni, and Rahmaddeni, "Perbandingan Algoritma Random Forest dan XGBoost untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien," *J. Ilm. BETRIK Besemah Teknol. Inf. dan Komput.*, vol. 15, no. 2, pp. 130–141, 2024, [Online]. Available: https://ejournal.pppmitpa.or.id/index.php/betrik/article/view/231

[11] Y. A. Rindri and A. Fitriyani, "Analisis Perbandingan Kinerja Algoritma Multilayer Perceptron dan K-Nearest Neighbor pada Klasifikasi Tipe Migrain Comparative Analysis of Multilayer Perceptron and K-Nearest Neighbor Algorithms in Migraine Type Classification," *J. Teknol. dan Inf.*, vol. 13, no. 1, pp. 44–55, 2023, doi: 10.34010/jati.v13i1.

[12] W. Ningsih, B. Alfianda, R. Rahmaddeni, and D. Wulandari, "Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 556–562, 2024, doi: 10.57152/malcom.v4i2.1253.

[13] Y. Septiawan and Chairani, "Perbandingan Akurasi Metode Deteksi Ujaran Kebencian dalam Postingan Twitter Menggunakan Metode SVM dan Decision Trees yang Dioptimalkan dengan Adaboost," *J. Tek.*, vol. 17, no. 2, pp. 297–299, 2023.

[14] L. G. R. Putra, D. D. Prasetya, and Mayadi, "Student Dropout Prediction Using Random Forest and XGBoost Method," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.

[15] Iddrus and D. W. Sari, "Penerapan Data Mining Menggunakan Algoritma Decision Tree C4 . 5 Untuk Memprediksi Mahasiswa Drop Out di Universitas Wiraraja," *J. Adv. Res. Inform.*, vol. 1, no. Juni, pp. 1–7, 2023.

[16] F. L. Hanis, U. Khaira, and D. Arsa, "Classification of Student Drop Out Risk Using Decision Tree C5 . 0 with Feature Selection Klasifikasi Status Mahasiswa Berisiko Drop Out Menggunakan Decision Tree C5 . 0 dengan Seleksi Fitur," vol. 5, no. October, pp. 1318–1330, 2025.

[17] A. R. B. Jamroni, W. Hadikristanto, and M. Fatchan, "Analisis Faktor dan Prediksi Atrisi untuk Optimalisasi Retensi Karyawan Menggunakan *Machine learning*," vol. 7, no. 3, 2025, doi: 10.32877/bt.v7i3.2301.

[18] D. N. Fahria, K. D. Tania, and R. D. Kurnia, "Analisis komparatif algoritma random forest, xgboost, dan catboost untuk klasifikasi tingkat stres pengguna media sosial 1)," *RABIT J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 1843–1853, 2026.

[19] M. Maida, M. A. Soeleman, H. P. Novitasari, and S. A. Rositasari, "Analisis Algoritma *Machine learning* dengan Teknik SMOTE untuk Peningkatan Sensitivitas Model Deteksi Sindrom Ovarium Polikistik (PCOS)," *RABIT J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 1013–1027, 2026.

[20] D. J. Yantono, N. R. Puspaningrum, and S. A. Zein, "Analisis Komparasi Algoritma *Machine learning* untuk Prediksi Harga Sewa Properti Komersial," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 6, no. January, pp. 149–155, 2026.

[21] E. W. Solang, F. X. Adu, A. Dharma, and N. Gunantara, "Evaluasi *Machine learning* untuk Prediksi Pembatalan Hotel dengan Threshold Adjustment dan Cost-Based Evaluation," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 6, no. January, pp. 193–204, 2026.

[22] W. A. Rayadhani and M. Rahardi, "Comparative Analysis of Random Forest , SVM , and Naive Bayes for Cardiovascular Disease Prediction," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3234–3243, 2025.

[23] A. A. Sitanggang, M. Y. Lase, S. P. Sipayung, T. Informatika, U. Katolik, and S. Thomas, "Prediksi Kelulusan Mahasiswa Menggunakan Logistic Regression dan Random Forest Berdasarkan Data Akademik," vol. 10, pp. 3503–3513, 2026.

[24] R. Mardianto, S. Quinevera, and S. Rochimah, "Perbandingan Metode Random Forest , Convolutional Neural Network , dan Support Vector Machine Untuk Klasifikasi Jenis Mangga," *J. Appl. Comput. Sci. Technol.*, vol. 5, no. 1, pp. 63–71, 2024.

[25] A. Wiraguna, S. Al Faraby, and Adiwijaya, "Klasifikasi Topik Multi Label pada Hadis Bukhari dalam Terjemahan Bahasa Indonesia Menggunakan Random Forest," *e-Proceeding Eng.*, vol. 6, no. 1, pp. 2144–2153, 2019.

[26] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.

[27] I. Hamzah, Muhammad Iqbal, and Rian Farta Wijaya, "Klasifikasi Jenis Penyakit dengan Algoritma Decision Tree Menggunakan Rapid Miner," *J. Nas. Teknol. Komput.*, vol. 4, no. 1, pp. 34–39, 2024, doi: 10.61306/jnastek.v4i1.127.

[28] A. M. Wahid, T. Turino, K. A. Nugroho, T. S. Maharani, D. Darmono, and F. S. Utomo, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis TF-IDF," *J. Pendidik. dan Teknol. Indones.*, vol. 4, no. 8, pp. 381–392, 2024.

[29] C. Bautista and D. Udjulawa, "Penerapan Random Forest Dan Borderline SMOTE Untuk Prediksi Risiko Drop Out Mahasiswa," *SINTECH J.*, vol. 8, no. 3, pp. 200–210, 2025.

[30] N. Khasanah, D. Uki, E. Saputri, F. Aziz, and T. Hidayat, "Studi Perbandingan Algoritma Random Forest dan K-Nearest Neighbors (KNN) dalam Klasifikasi Gangguan Tidur," *Comput. Sci.*, vol. 5, no. 1, pp. 17–25, 2025.

[31] E. Sahelvi, P. Cikita, and R. M. Sapitri, "Perbandingan Algoritma K-Nearest Neighbors dan Random Forest untuk Rekomendasi Gaya Hidup Sehat dalam Mencegah Penyakit Jan," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. July, pp. 830–840, 2025.

[32] Samuel, Idmi, and G. Triyono, "Analisis perbandingan model naïve bayes dan c4.5 untuk prediksi stroke berdasarkan riwayat data medis dengan pendekatan matriks korelasi," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 4, pp. 3749–3759, 2025.

[33] N. H. Setyawan and N. Wakhidah, "Analisis Perbandingan Metode Logistic Regression, Random Forest, Gradient Boosting Untuk Prediksi Diabetes," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 1, pp. 150–162, 2025, doi: 10.29100/jipi.v10i1.5743.

[34] A. M. Wahid, Turino, K. A. Nugroho, D. Titi Safitri4, and F. S. Utomo, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis Hyperparameter Optimization," *J. Pendidik. dan Teknol. Indones.*, vol. 4, no. January, pp. 381–392, 2025.

[35] D. A. Ardhani and K. D. Tania, "Knowledge Discovery on E-Commerce Customer Churn Using Interpretable *Machine learning*: A Comparative Study of SHAP-Based Classifiers," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2695–2702, 2025, doi: 10.30871/jaic.v9i5.10811.

[36] A. P. Istiwana, R. R. Sani, and Y. T. C. Pramudi, "Pendekatan Explainable *Machine learning* untuk Analisis Faktor Drop Out Mahasiswa Menggunakan XGBoost," *J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 1074–1083, 2026.

[37] L. N. Hapsari, I. Fannani, Y. Rahmawati, and A. Muhariya, "Explainable *Machine learning* untuk Prediksi Risiko Penyakit Jantung Menggunakan Random Forest dan Analisis SHAP," *REMIK Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 10, no. 1, pp. 190–199, 2026, doi: http://doi.org/10.33395/remik.v10i1.15766.

[38] A. A. Sutrisno *et al.*, "Klasifikasi Risiko Gagal Bayar Kredit Menggunakan XGboost Dengan Model Explainable Berbasis SHAP," vol. 10, no. 1, pp. 296–301, 2026.

[39] D. R. B. Lubis and A. Armansyah, "Klasifikasi Mahasiswa Berpotensi Dropout Menggunakan Metode Regresi Logistik," *Techno.Com*, vol. 24, no. 3, pp. 768–778, 2025, doi: 10.62411/tc.v24i3.13526.

[40] R. Syahranita, S. Suhartono, and S. Zaman, "Prediksi Kategori Kelulusan Mahasiswa Menggunakan Metode Regresi Logistik Multinomial," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 8, no. 2, pp. 102–111, 2023, doi: 10.14421/jiska.2023.8.2.102-111.

[41] S. A. Hendrawan, "Analysis of the Digital System-Based Academic Factors and the Risk of Student Dropout Using Chi-Square Test for Data Driven-Interventions," *Edu Cendikia J. Ilm. Kependidikan*, vol. 5, no. 02, pp. 602–616, 2025, doi: 10.47709/educendikia.v5i02.6736.

[42] A. Kusumawardani and D. R. Hidayat, "A Literature Review on Risk and Protection Factors of Burnout in Healthcare Students," *Psikologika J. Pemikir. dan Penelit. Psikol.*, vol. 28, no. 2, pp. 243–262, 2023, doi: 10.20885/psikologika.vol28.iss2.art6.

[43] R. P. Nugraha, G. F. Laxmi, and F. Riana, "Penerapan K-Means ++ Untuk Pengelompokan Mahasiswa Berpotensi Drop Out (Studi Kasus : Universitas Ibn Khaldun Bogor)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3493–3500, 2024.

[44] A. S. B. Lomi, A. A. Pekuwali, and R. T. Abineno, "Pengelompokan Mahasiswa Berpotensi Drop Out pada Program Studi Teknik Infromatika Menggunakan Metode K-Means Clustering," *SATI Sustain. Agric. Technol. Innov.*, no. 1, p. 340, 2024, [Online]. Available: https://www.ojs.unkriswina.ac.id/index.php/semnas-FST/article/view/935

[45] S. Fahsadila, "The Effect of Support Programs on Academic Resilience in Dropout Students: A Systematic Literature Review," *J. BIKOTETIK (Bimbingan dan Konseling Teor. dan Prakt.*, vol. 09, no. 1, pp. 143–155, 2025.

[46] W. Laucu, "The Impact of Part-Time Work on Student Academic Achievement," *Contin. Indones. J. Islam. Community Dev.*, vol. 2, pp. 65–79, 2023.