

Pengembangan Model Prediksi Kelulusan Siswa SMA ke Perguruan Tinggi Negeri (PTN) Menggunakan *Machine learning* Berbasis CRISP-DM

Achmad Arif Mulyana^{1*}, Samidi^{1,2}

¹ Universitas Padjadjaran, Indonesia

² Universitas Budi Luhur, Indonesia

 achmad12002@mail.unpad.ac.id*

ABSTRACT

The development of educational technology has witnessed the use of *machine learning* as a basis for data-driven decision making in the education sector. The objective of this research is to build a predictive model of admission of high school students at State Universities (PTN) using a *machine learning* method based on the Cross Industry Standard Process for Data Mining (CRISP-DM) framework. The data was obtained from the Student Center tutoring institution and consisted of academic, non-academic, and socio-economic attributes. Three classification algorithms i.e. Decision Tree, Random Forest, *Naive Bayes* were used. The split percentage method was employed in the model evaluation with the training and testing data division schemes of 70:30, 80:20, and 90:10. The results indicated that the Random Forest algorithm had the highest average *accuracy* at 96.63%, followed by the Decision Tree with 95.48% and the *Naive Bayes* with 94.06%. It was found that class ranking, try-out scores, school accreditation and attendance rate are several variables that significantly affect the students' chances to be accepted in PTN. These findings demonstrate that academic success is determined not only by learning achievement, but also by study discipline and educational environment quality. This research contributes to the advancement of educational data mining and supports educational innovation by using *machine learning* as a decision support system for student guidance and academic evaluation.

Keywords: CRISP-DM, *Decision Tree*, *machine learning*, *Naive Bayes*

ARTICLE INFO

Article history:

Received

February 26, 2026

Revised

June 04, 2026

Accepted

June 19, 2026

Journal Homepage

<https://attractivejournal.com/index.php/aj/>

This is an open access article under the CC BY SA license

<https://creativecommons.org/licenses/by-sa/4.0/>

@ 2026 by the authors

Published by

CV. Creative Tugu Pena

PENDAHULUAN

Pendidikan pada jenjang Sekolah Menengah Atas (SMA) memegang peranan penting dalam membentuk sumber daya manusia yang berkualitas dan mampu bersaing. Salah satu tolok ukur keberhasilan pada peran ini adalah keberhasilan siswa dalam melanjutkan studi ke Perguruan Tinggi Negeri (PTN) dimana PTN selama ini dipandang memiliki mutu pendidikan yang lebih baik dengan biaya yang relatif terjangkau. Namun, tingginya minat siswa untuk masuk PTN tidak diimbangi dengan ketersediaan daya tampung yang memadai sehingga persaingan menjadi semakin ketat dan tingkat kelulusan cenderung rendah setiap tahunnya. Kondisi tersebut menunjukkan perlunya upaya pembinaan yang lebih terarah, terencana, dan terukur dengan pemanfaatan data agar peluang keberhasilan siswa dapat ditingkatkan secara optimal (AlOtaibi, 2025; Andrianof et al., 2025).

Proses pembinaan siswa di sekolah maupun lembaga bimbingan belajar saat ini masih mengandalkan pendekatan tradisional yang berfokus pada aspek akademik saja, seperti nilai rapor dan hasil *try out*. Padahal keberhasilan siswa dalam seleksi masuk PTN tidak hanya ditentukan oleh kemampuan akademik, tetapi juga dipengaruhi oleh berbagai faktor lain,

seperti motivasi belajar, kedisiplinan, keaktifan dalam organisasi, tingkat kehadiran, serta kondisi sosial ekonomi keluarga (AlOtaibi, 2025; Arifin et al., 2026; Sun et al., 2023). Jika proses pembinaan ini hanya terfokus pada kemampuan akademik maka berisiko menyebabkan potensi siswa tidak teridentifikasi secara menyeluruh (Hussain & Khan, 2023; Thakur & Rawat, 2025). Akibatnya, strategi pembinaan yang nantinya diberikan akan kurang tepat sasaran dan tidak mampu mengoptimalkan peluang keberhasilan siswa secara maksimal.

Perkembangan teknologi informasi saat ini membuka peluang untuk menerapkan pendekatan yang lebih modern dan berbasis data dalam dunia pendidikan. Salah satu pendekatan berbasis data yang banyak dimanfaatkan adalah *machine learning*, yang merupakan bagian dari kecerdasan buatan yang mampu mengenali pola dari data historis dan menghasilkan prediksi secara otomatis dengan tingkat akurasi yang tinggi (Guanin-Fajardo et al., 2024; Okoye et al., 2024). Pendekatan *machine learning* telah banyak digunakan untuk menganalisis dan memprediksi performa siswa dengan mempertimbangkan berbagai variabel secara simultan pada bidang pendidikan (Airlangga, 2024; Issah et al., 2023; Wang & Yu, 2025). Beberapa penelitian terbaru menunjukkan bahwa algoritma seperti Decision Tree dan Random Forest tidak hanya memiliki tingkat akurasi yang tinggi, tetapi juga mudah diinterpretasikan sehingga relevan untuk digunakan dalam pengambilan keputusan di bidang Pendidikan (E. Ahmed, 2024; AlOtaibi, 2025; Andrianof et al., 2025). Oleh karena itu, pemanfaatan *machine learning* dalam memprediksi kelulusan siswa menjadi alternatif solusi yang potensial untuk meningkatkan kualitas pembinaan berbasis data.

Pentingnya penelitian ini terletak pada beberapa permasalahan mendasar yang belum terselesaikan dalam sistem pendidikan Indonesia. Pertama, lembaga bimbingan belajar selama ini belum memiliki sistem prediksi yang komprehensif dan berbasis data untuk mengidentifikasi siswa yang berpotensi lulus PTN sejak dini. Kedua, pendekatan konvensional yang bersifat reaktif dan subjektif perlu digantikan dengan pendekatan yang lebih sistematis, objektif, dan terukur. Ketiga, integrasi faktor akademik, non-akademik, dan sosial-ekonomi dalam satu model prediksi masih sangat terbatas dalam konteks pendidikan Indonesia. Penelitian ini hadir untuk menjawab celah tersebut dengan mengembangkan model prediksi berbasis *machine learning* yang komprehensif, terstruktur, dan relevan dengan konteks lokal.

Meskipun demikian, penerapan *machine learning* untuk memprediksi kelulusan siswa ke perguruan tinggi, khususnya di Indonesia, masih belum banyak dilakukan. Sebagian besar penelitian yang ada, seperti yang dilakukan oleh Issah et al. (2023) dan Kaensar & Wongnin (2023), lebih berfokus pada prediksi prestasi akademik secara umum tanpa mempertimbangkan konteks seleksi masuk PTN yang sangat spesifik. Penelitian Guanin-Fajardo et al. (2024) menggunakan teknik *machine learning* untuk memprediksi keberhasilan akademik di tingkat perguruan tinggi, namun tidak mempertimbangkan faktor non-akademik dan sosial-ekonomi secara bersamaan. Sementara penelitian Azizah et al. (2024) menggunakan *blended learning* sebagai konteks prediksi, berbeda dari konteks bimbingan belajar pra-PTN yang dikaji dalam penelitian ini. Demikian pula, penelitian Nimy et al. (2023) berfokus pada identifikasi siswa berisiko di perguruan tinggi, bukan pada prediksi kelulusan seleksi masuk. Hal ini menunjukkan adanya celah penelitian (*research gap*) yang signifikan dan perlu dikaji lebih lanjut melalui pengembangan model prediksi yang lebih komprehensif, kontekstual, dan terintegrasi.

Berdasarkan permasalahan tersebut, penelitian ini difokuskan pada pengembangan model prediksi kelulusan siswa SMA ke PTN dengan memanfaatkan pendekatan *machine learning* yang menggabungkan berbagai variabel akademik, non-akademik, dan sosial ekonomi. Proses pengembangan model dilakukan dengan menggunakan metodologi CRISP-DM, yang mencakup tahapan pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi, hingga implementasi (Schröer et al., 2021). Dalam penelitian ini, beberapa algoritma klasifikasi seperti *Decision Tree*, *Random Forest*, dan *Naive Bayes* digunakan dan dibandingkan untuk mengetahui model yang memiliki kinerja terbaik dalam memprediksi peluang kelulusan siswa.

Tujuan penelitian ini adalah: (1) mengembangkan model prediksi kelulusan siswa SMA ke PTN dengan mengintegrasikan variabel akademik, non-akademik, dan sosial-ekonomi; (2) membandingkan kinerja algoritma *Decision Tree*, *Random Forest*, dan *Naive Bayes* dalam konteks

prediksi kelulusan PTN; (3) mengidentifikasi variabel-variabel yang paling berpengaruh terhadap peluang kelulusan siswa; dan (4) memberikan rekomendasi bagi lembaga bimbingan belajar dalam menyusun strategi pembinaan berbasis data. Hasil dari penelitian ini diharapkan tidak hanya memberikan kontribusi dalam pengembangan ilmu pengetahuan, tetapi juga memberikan manfaat praktis sebagai sistem pendukung keputusan bagi lembaga pendidikan.

Penerapan *machine learning* dalam sektor pendidikan menjadi bagian penting dari transformasi digital yang memperkuat praktik evaluasi dan pengambilan keputusan berbasis data. Dengan memanfaatkan data historis siswa, model prediktif dapat digunakan untuk menganalisis pola capaian akademik, perilaku belajar, serta faktor-faktor lain yang berhubungan dengan peluang keberhasilan siswa (Almalawi et al., 2024; Diponegoro et al., 2021; Jang et al., 2022). Berbeda dari pendekatan konvensional yang cenderung bersifat manual dan reaktif, pendekatan berbasis *machine learning* memungkinkan lembaga pendidikan melakukan pemetaan risiko dan potensi siswa secara lebih sistematis, objektif, dan cepat (W. Ahmed et al., 2025; de Oliveira et al., 2024; Okoye et al., 2024). Dalam konteks prediksi kelulusan PTN, model yang dikembangkan dapat membantu mengelompokkan siswa berdasarkan tingkat peluang keberhasilannya, sehingga strategi pembinaan dapat disusun secara lebih personal dan sesuai dengan kebutuhan masing-masing siswa (Abuzinadah et al., 2023; Azizah et al., 2024; Nimy et al., 2023).

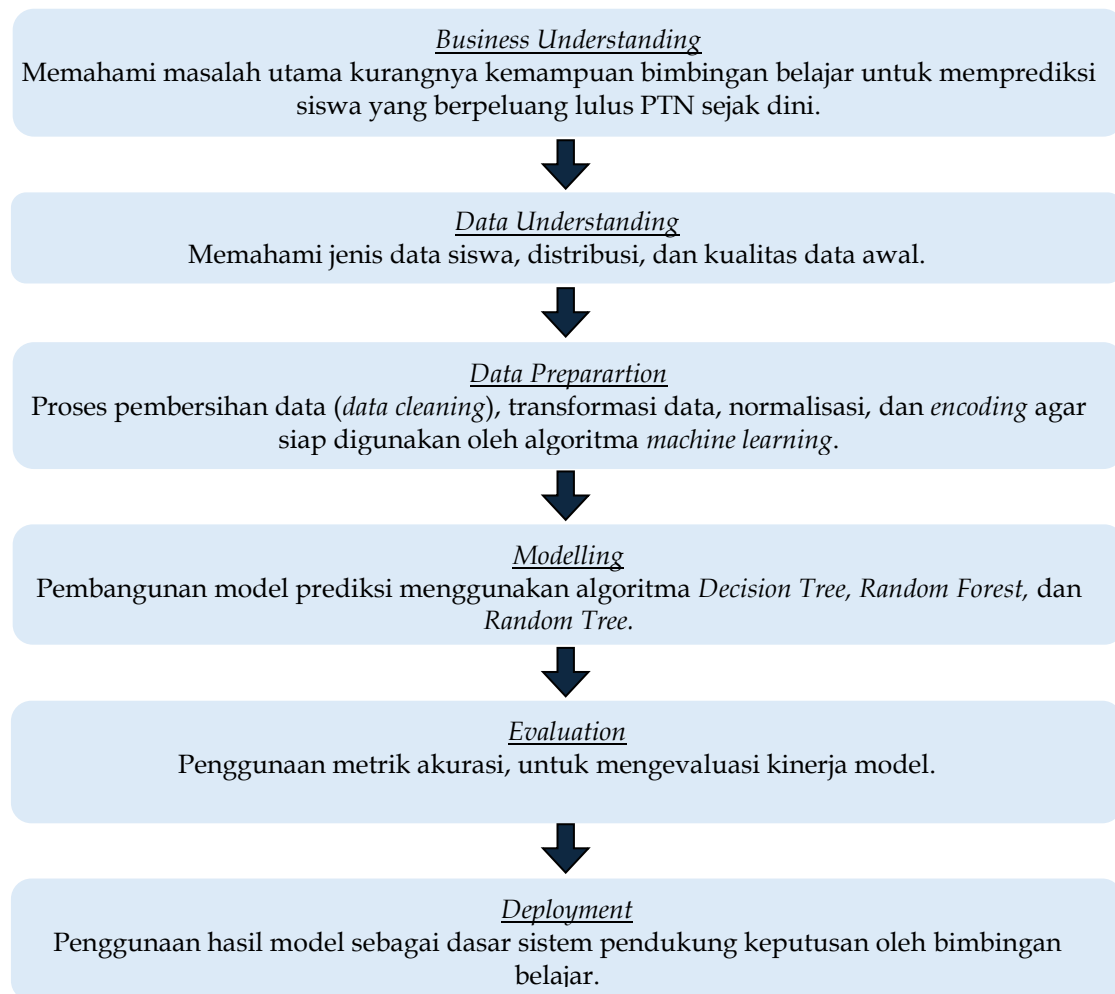
Dalam konteks *educational data mining* (EDM), berbagai teknik telah dikembangkan untuk menganalisis data pendidikan guna mendukung pengambilan keputusan yang lebih baik. Romero & Ventura (2020) mendefinisikan EDM sebagai bidang ilmu yang berfokus pada pengembangan metode untuk mengeksplorasi data yang berasal dari lingkungan pendidikan dan menggunakan temuan tersebut untuk pemahaman siswa dan konteks pembelajaran yang lebih baik. Dutt et al. (2017) dalam tinjauan sistematis mereka mengidentifikasi bahwa teknik klasifikasi, termasuk *Decision Tree*, *Random Forest*, dan *Naive Bayes*, merupakan teknik yang paling banyak digunakan dalam EDM untuk prediksi kinerja siswa. Chen et al. (2025) juga menegaskan bahwa penerapan *machine learning* dalam Pendidikan tinggi telah memberikan kontribusi signifikan dalam meningkatkan prediksi performa akademik mahasiswa, keterlibatan belajar (*learning engagement*), serta kemampuan mengidentifikasi mahasiswa yang berisiko mengalami penurunan prestasi akademik. Melalui berbagai algoritma seperti *Decision Tree*, *Random Forest*, *Support Vector Machine*, dan *Neural Network*, *machine learning* mampu mendukung proses intervensi dini dan pengambilan keputusan yang lebih efektif dalam lingkungan pendidikan. Penelitian ini turut memperkaya kajian *Educational Data Mining* (EDM) di Indonesia dengan memberikan bukti empiris dari konteks bimbingan belajar pra-perguruan tinggi yang masih relatif terbatas dalam literatur penelitian sebelumnya.

Secara teoritis, kerangka kerja yang digunakan dalam penelitian ini dibangun di atas landasan tiga teori utama. Pertama, teori *predictive analytics* yang menyatakan bahwa pola historis dapat digunakan untuk memproyeksikan kejadian masa depan (Waheed et al., 2020). Kedua, teori belajar konstruktivistik yang menekankan bahwa proses belajar bersifat aktif dan dipengaruhi oleh berbagai faktor internal maupun eksternal siswa (Kotsiantis, 2012). Ketiga, teori modal sosial dan kapital budaya Bourdieu yang menjelaskan bagaimana latar belakang sosial-ekonomi dapat memengaruhi akses dan peluang Pendidikan (Sultana et al., 2017). Integrasi tiga landasan teoritis ini memperkuat relevansi pemilihan variabel akademik, non-akademik, dan sosial-ekonomi dalam model prediksi yang dikembangkan dalam penelitian ini.

METODE

Metode pendekatan yang bersifat sistematis, fleksibel, dan sesuai untuk pengembangan model prediksi berbasis *machine learning* dalam penelitian ini adalah CRISP-DM (*Cross Industry Standard Process for Data Mining*) (Schröer et al., 2021). CRISP-DM merupakan kerangka kerja yang banyak digunakan dalam penelitian *data mining* dan *machine learning* karena menyediakan panduan terstruktur mulai dari pemahaman masalah bisnis hingga implementasi model. Guanin-Fajardo et al. (2024) menegaskan bahwa penggunaan CRISP-DM dalam konteks pendidikan meningkatkan sistematisasi proses pemodelan dan menghasilkan pengetahuan

yang lebih valid secara ilmiah. Tahapan metode CRISP-DM dalam penelitian ini meliputi enam fase utama sebagai berikut.



Gambar 1. Tahapan Metode CRISP-DM

Penelitian ini menggunakan data sekunder yang diperoleh dari sistem administrasi Pusat Bimbingan Belajar Student Center. Tabel 1 menampilkan daftar atribut yang digunakan secara lengkap. Data mencakup seluruh siswa SMA yang terdaftar dan telah mengikuti seleksi masuk PTN selama periode tahun 2021 hingga 2025, dengan total sebanyak 2.052 data siswa. Data digunakan seluruhnya sebagai sampel penelitian. Hal ini dilakukan karena jumlah data termasuk dalam rentang yang cukup untuk pengembangan model *machine learning*, sehingga tidak diperlukan pengambilan *sampling*. Atribut yang digunakan dalam pemodelan terdiri atas 12 variabel yang merepresentasikan tiga aspek utama, yaitu akademik, non-akademik, dan sosial-ekonomi.

Komponen variabel akademik adalah hasil ujian, posisi peringkat, dan catatan kehadiran, yang semuanya merupakan bagian dari catatan kegiatan panduan belajar. Komponen variabel non-akademik adalah peran siswa selama kegiatan panduan belajar dan keanggotaan dalam organisasi. Komponen variabel sosial ekonomi adalah yang berasal dari biodata orang tua, termasuk tingkat pendidikan, pekerjaan, pendapatan, jumlah anak, dan tanggungan. Dalam studi ini, variabel tambahan juga dimasukkan seperti riwayat partisipasi dalam panduan belajar, status pembayaran, dan riwayat posisi peringkat untuk memperkaya analisis tentang prospek keberhasilan siswa dalam mengakses pendidikan ke PTN.

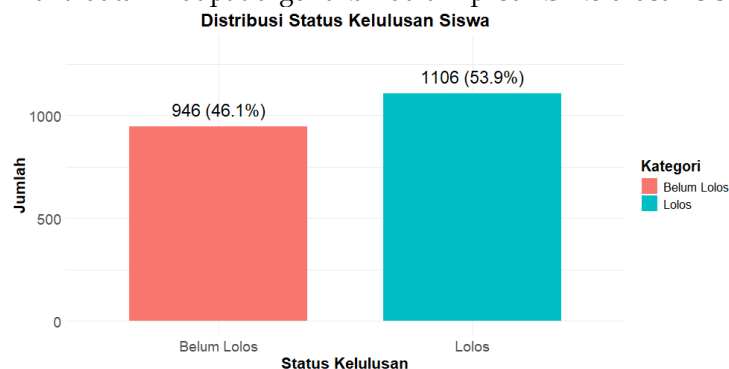
Tabel 1. Keterangan Atribut yang Digunakan

No	Nama Atribut	Kelompok Atribut	Jenis Data	Keterangan
1	Akreditasi Sekolah	Akademik	Kategorik	A/B
2	Jurusan Sekolah	Akademik	Kategorik	IPA/IPS
3	Nilai Tryout	Akademik	Kategorik	Di Bawah 500/500-559/Di Atas 600
4	Peringkat Kelas	Akademik	Kategorik	10 Besar/20 Besar/Di Atas 20 Besar
5	Riwayat Mengikuti Bimbel	Akademik	Kategorik	Pernah/Tidak Pernah
6	Kehadiran	Non-Akademik	Kategorik	Cukup/Rajin/Tidak Rajin
7	Keaktifan Organisasi	Non-Akademik	Kategorik	Iya/Tidak
8	Pendidikan Terakhir Ayah	Sosial-Ekonomi	Kategorik	SMP Atau Lainnya/SMA Atau SMK/S1
9	Pendidikan Terakhir Ibu	Sosial-Ekonomi	Kategorik	SMP Atau Lainnya/SMA Atau SMK/S2
10	Penghasilan Orang Tua	Sosial-Ekonomi	Kategorik	Kurang Dari 2 Juta/2-4 Juta/Lebih Dari 6 Juta
11	Jumlah Tanggungan	Sosial-Ekonomi	Numerik	1-6
12	Status Pembayaran Bimbel	Sosial-Ekonomi	Kategorik	Lancar/Tidak Lancar
13	Status Kelulusan Status Lolos PTN	Sosial-Ekonomi	Kategorik	Lolos/Tidak Lolos

HASIL DAN PEMBAHASAN

Dalam penelitian ini, data yang digunakan berkaitan dengan tujuan analisis yang dilakukan yaitu prediksi kelulusan siswa SMA ke PTN. Data terdiri dari beberapa atribut yang dikelompokkan ke dalam tiga kategori utama, yaitu atribut akademik, non-akademik, dan sosial-ekonomi. Atribut target pada penelitian ini yaitu status kelulusan siswa ke PTN. Setiap *record* dalam dataset ini merepresentasikan satu individu siswa dengan sejumlah atribut yang menggambarkan karakteristik dan latar belakang siswa secara komprehensif, sehingga kita dapat mengidentifikasi faktor-faktor yang memengaruhi keberhasilan siswa dalam seleksi masuk PTN.

Untuk memberikan gambaran awal tentang distribusi data penelitian, visualisasi proporsi terhadap status kelulusan siswa ke PTN perlu dilakukan untuk dapat memahami karakteristik dataset yang digunakan dalam proses klasifikasi. Hal ini dapat membantu memahami bagaimana data ini dapat digunakan dalam prediksi kelulusan siswa SMA ke PTN.

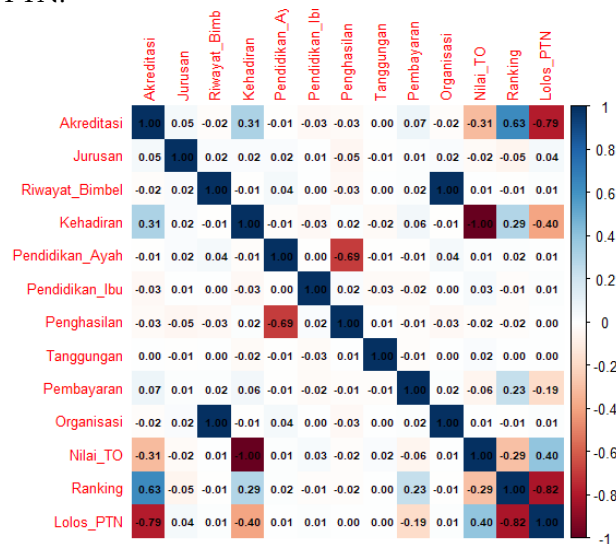


Gambar 2. Distribusi Status Kelulusan Siswa ke PTN

Berdasarkan visualisasi tersebut, distribusi data menunjukkan bahwa sebanyak 1.106 siswa (53,9%) yang berhasil lolos PTN, sementara 946 siswa (46%) yang belum lolos. Distribusi yang relatif seimbang ini menunjukkan bahwa dataset yang digunakan dalam penelitian ini bebas dari masalah ketidakseimbangan kelas (*class imbalance*) yang parah, yang sering menjadi tantangan dalam penelitian EDM (Nimy et al., 2023). Hal ini juga berkontribusi pada tingginya

akurasi model yang dicapai, karena model tidak perlu menangani bias akibat ketidakseimbangan distribusi kelas. Temuan ini konsisten dengan penelitian Zeineddine et al. (2021) yang menunjukkan bahwa distribusi kelas yang seimbang dapat meningkatkan performa model prediksi secara signifikan.

Selanjutnya, tahap awal berupa *business understanding* dilakukan identifikasi kebutuhan lembaga bimbingan belajar dalam memprediksi peluang kelulusan siswa ke PTN sejak dini. Kemudian, dilakukan eksplorasi karakteristik atribut akademik, non-akademik, dan sosial-ekonomi siswa. Pada tahap *data cleaning*, dilakukan penanganan untuk data yang hilang, duplikasi data, serta inkonsistensi nilai atribut. Tujuan dari proses ini adalah untuk meningkatkan kualitas data sehingga model yang dihasilkan memiliki tingkat akurasi dan reliabilitas yang lebih baik. Selain itu dilakukan proses *pre-processing* untuk memastikan bahwa data siap digunakan dalam proses pembuatan model machine learning. Matriks korelasi dibentuk untuk mengetahui hubungan antar variabel yang digunakan dalam penelitian. Visualisasi ini membantu mengidentifikasi pola hubungan antara variabel akademik, non-akademik, sosial-ekonomi, dan status kelulusan siswa. Dengan demikian, kita dapat memahami bagaimana variabel-variabel ini berhubungan dan mempengaruhi peluang kelulusan siswa ke PTN.



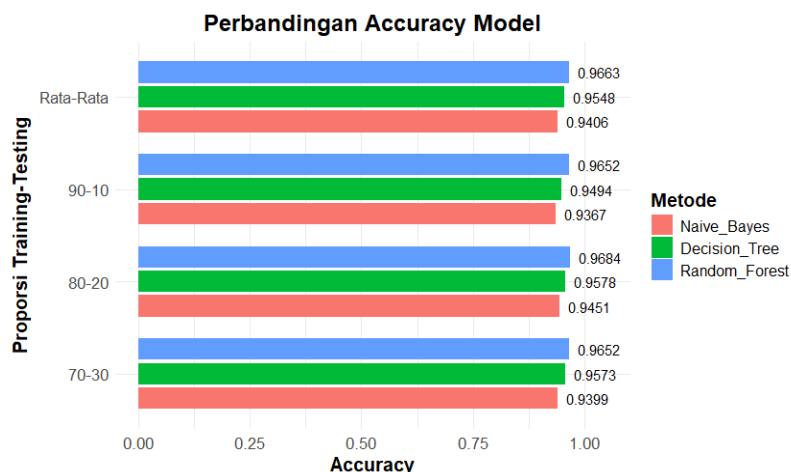
Gambar 3. Heatmap Korelasi Antarvariabel

Hasil analisis korelasi dengan *heatmap* menunjukkan bahwa variabel Ranking tercatat memiliki keterkaitan paling erat dengan status kelulusan siswa ke PTN, dengan nilai korelasi sekitar -0,82. Hal Ini berarti bahwa siswa dengan peringkat yang lebih tinggi atau nomor urut yang lebih kecil cenderung memiliki peluang lebih besar untuk diterima di PTN. Selain itu, variabel Nilai *Try Out* juga menunjukkan hubungan yang cukup signifikan terhadap kelulusan PTN, yang menunjukkan bahwa tingkat kesiapan akademik siswa sangat penting dalam menentukan keberhasilan mereka pada proses seleksi masuk PTN. Kehadiran juga menunjukkan korelasi positif, meskipun nilainya tergolong rendah, yang berarti kedisiplinan siswa dalam mengikuti kegiatan belajar tetap memberikan kontribusi terhadap peluang keberhasilan mereka. Visualisasi *heatmap* ini menunjukkan bahwa faktor-faktor akademik, khususnya peringkat kelas dan nilai *try out*, merupakan prediktor yang paling dominan dalam menentukan kelulusan siswa ke PTN bila dibandingkan dengan faktor-faktor sosial-ekonomi.

Selanjutnya pada tahap *modeling* akan menggunakan algoritma klasifikasi *Naive Bayes*, *Decision Tree*, dan *Random Forest*. Ketiga algoritma ini dipilih karena karakteristiknya yang berbeda dalam melakukan klasifikasi data. *Naive Bayes* menggunakan pendekatan probabilistik, *Decision Tree* membentuk struktur pohon keputusan, sedangkan *Random Forest* mengombinasikan banyak *decision tree* untuk meningkatkan performa prediksi dan mengurangi *overfitting*. Kita akan menggunakan metode *split percentage* dengan tiga skenario pembagian

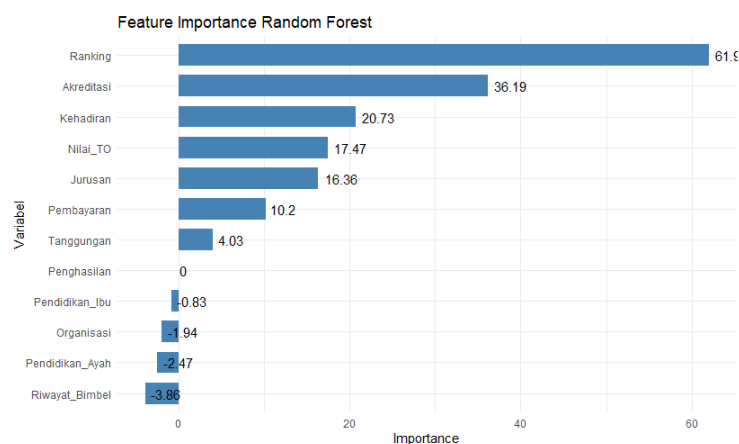
data *training* dan *testing*, yaitu 70:30, 80:20, dan 90:10. Model akan dievaluasi menggunakan nilai *accuracy* untuk mengukur kemampuan model dalam mengklasifikasikan siswa ke dalam kategori lulus dan tidak lulus PTN.

Tabel 2. Perbandingan Akurasi pada Split Percentage Method



Gambar 4. Perbandingan Akurasi Model

Perbandingan akurasi model pada Gambar 4 menunjukkan bahwa algoritma *Random Forest* menunjukkan performa terbaik dengan rata-rata akurasi sebesar 96,63%. Kemudian, diikuti oleh *Decision Tree* dengan akurasi sebesar 95,48%, dan *Naive Bayes* dengan akurasi sebesar 94,06%. *Random Forest* juga menunjukkan performa yang relatif stabil dalam berbagai skenario pembagian data *training* dan *testing*. Ini menunjukkan bahwa *Random Forest* dapat menangkap pola hubungan kompleks antara faktor akademik, non-akademik, dan sosial-ekonomi dalam memprediksi kelulusan siswa ke perguruan tinggi negeri. Pendekatan *ensemble learning* pada *Random Forest* memungkinkan model untuk menangkap hubungan *non-linear* antar variabel lebih baik dibandingkan dengan *Naive Bayes* yang menggunakan asumsi independensi antar atribut.

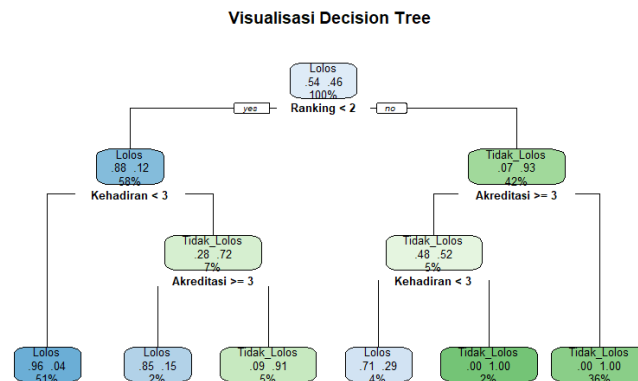


Gambar 5. Feature Importance pada Algoritma Random Forest

Grafik *Feature Importance Random Forest* pada Gambar 5 menunjukkan bahwa variabel *Ranking* menjadi faktor paling dominan dalam memprediksi kelulusan siswa ke perguruan tinggi negeri. Variabel *Akreditasi* sekolah, *Kehadiran*, dan *Nilai Try Out* juga memberikan kontribusi besar terhadap hasil prediksi. Temuan ini menunjukkan bahwa faktor akademik dan kedisiplinan belajar memiliki pengaruh utama terhadap peluang siswa lolos perguruan tinggi negeri. Sementara itu, penghasilan dan pendidikan orang tua yang termasuk dalam faktor sosial-ekonomi memiliki pengaruh yang relatif lebih rendah dalam model prediksi. Temuan ini sesuai dengan konsep pendidikan integratif yang menekankan pentingnya keseimbangan antara aspek kognitif dan karakter dalam proses pembelajaran. Penelitian ini tidak hanya

memberikan kontribusi pada pengembangan model prediksi berbasis *machine learning*, tetapi juga mendukung pengembangan pendekatan pendidikan berbasis data dalam proses pembinaan siswa.

Sementara itu, algoritma *Decision Tree* juga menunjukkan performa yang cukup baik karena mampu membentuk aturan klasifikasi. Model ini unggul pada aspek interpretabilitas karena pola keputusan yang dihasilkan dapat dipahami dengan lebih mudah oleh pengguna non-teknis seperti guru atau pengelola bimbingan belajar. Gambaran pola klasifikasi yang terbentuk menggunakan *Decision Tree* dapat dilihat pada gambar yang disediakan..



Gambar 6. Struktur *Decision Tree* Prediksi Kelulusan Siswa ke PTN

Kelulusan siswa ke perguruan tinggi negeri dipengaruhi oleh beberapa faktor. Faktor utama adalah ranking siswa di kelas. Siswa dengan ranking tinggi memiliki peluang lebih besar untuk lolos PTN. Selain itu, kehadiran dan akreditasi sekolah juga mempengaruhi kelulusan. Siswa dengan kehadiran yang baik dan berasal dari sekolah dengan akreditasi yang baik memiliki peluang lebih besar untuk lolos PTN.

Kemudian, algoritma *Naive Bayes* masih mampu memberikan hasil yang cukup baik meskipun memiliki tingkat akurasi yang lebih rendah dibandingkan dengan dua model sebelumnya. Ini bisa terjadi karena asumsi bahwa variabel-variabel dalam data pendidikan tidak saling bergantung, yang sebenarnya tidak tepat. Data pendidikan memiliki faktor-faktor seperti akademik, non-akademik, dan sosial-ekonomi yang saling terkait secara kompleks. Karena asumsi ini tidak terpenuhi, kemampuan model dalam memahami hubungan antarvariabel menjadi terbatas, yang pada akhirnya mempengaruhi akurasi. Meskipun demikian, algoritma ini masih bisa digunakan sebagai pembanding untuk mengevaluasi kinerja model lain yang lebih kompleks dalam memprediksi kelulusan siswa ke perguruan tinggi.

Temuan penelitian ini sejalan dengan studi yang dilakukan oleh Thakur & Rawat (2025) yang menunjukkan bahwa algoritma machine learning mampu menghasilkan prediksi performa akademik dengan tingkat akurasi yang tinggi melalui pemanfaatan atribut akademik dan perilaku belajar siswa. Hasil tersebut memperkuat temuan bahwa pendekatan berbasis data dapat digunakan sebagai sistem pendukung keputusan dalam lingkungan pendidikan. Hasil penelitian ini juga konsisten dengan penelitian Caday et al. (2025) yang menunjukkan bahwa model prediktif berbasis *Decision Tree*, *Naive Bayes*, dan *Random Forest* dapat digunakan sebagai sistem pendukung keputusan untuk membantu institusi pendidikan mengidentifikasi peluang keberhasilan akademik siswa sejak dini.

Untuk memahami lebih dalam mengapa *Random Forest* mengungguli algoritma lainnya dalam konteks penelitian ini, perlu dipahami mekanisme kerjanya secara lebih mendalam. *Random Forest* merupakan metode *ensemble learning* yang membangun banyak pohon keputusan selama pelatihan dan menghasilkan kelas yang merupakan modus dari kelas-kelas pohon individual (Breiman, 2001). Setiap pohon dalam *Random Forest* dilatih pada subset acak dari data pelatihan dan subset acak dari fitur, yang mengurangi korelasi antar pohon dan meningkatkan diversitas model. Dalam konteks data pendidikan yang cenderung memiliki hubungan non-linear antar variabel, kemampuan *Random Forest* untuk menangkap kompleksitas hubungan tersebut menjadi keunggulan yang signifikan. Hasil penelitian ini

menunjukkan bahwa algoritma *Random Forest* memberikan kinerja terbaik, yang sejalan dengan penelitian Andrianof yang menyatakan bahwa *Random Forest* efektif dalam mengelola data akademik dengan pola hubungan yang rumit. Kelebihan dari *ensemble learning* dalam *Random Forest* memungkinkan model ini mencapai akurasi yang lebih konsisten dibandingkan dengan metode klasifikasi lainnya. Temuan serupa juga diperoleh oleh Matzavela & Alepis (2021), yang menunjukkan bahwa *Decision Tree* mampu mengidentifikasi atribut akademik yang paling berpengaruh terhadap keberhasilan siswa sekaligus menghasilkan model yang mudah diinterpretasikan oleh pengambil keputusan pendidikan.

Hasil ini juga mendukung temuan Issah et al. (2023) yang menemukan bahwa variabel akademik seperti nilai, prestasi sebelumnya, dan capaian pembelajaran merupakan prediktor yang paling sering muncul dan paling signifikan dalam berbagai model *machine learning* untuk prediksi performa akademik. Penelitian Arifin et al. (2026) juga menunjukkan bahwa menggabungkan faktor akademik, sosial-ekonomi, dan perilaku siswa dapat memperbaiki kualitas model klasifikasi kinerja akademik. Namun, penelitian ini menemukan bahwa faktor sosial-ekonomi memiliki peran yang lebih kecil dibandingkan faktor akademik dan disiplin belajar dalam memprediksi kelulusan siswa ke PTN.

Studi ini juga sejalan dengan penelitian Guanin-Fajardo et al. (2024) yang menyatakan bahwa metode *machine learning* dapat membantu lembaga pendidikan mengidentifikasi siswa yang berisiko menghadapi kesulitan akademik lebih awal. Dengan demikian, model prediksi yang dihasilkan tidak hanya berguna sebagai alat klasifikasi, tetapi juga sebagai dasar untuk pengambilan keputusan dalam merumuskan strategi pembinaan siswa yang lebih efektif. Penerapan CRISP-DM dalam penelitian ini mengikuti rekomendasi Schröer et al. (2021) yang menyatakan bahwa CRISP-DM masih menjadi kerangka kerja yang paling banyak digunakan dalam proyek data mining karena menyediakan alur yang sistematis mulai dari *business understanding* hingga *deployment*. Pendekatan ini membantu memastikan bahwa pengembangan model *machine learning* dilakukan secara terstruktur dan dapat direplikasi pada konteks pendidikan lainnya.

Meskipun model yang digunakan pada penelitian ini cukup akurat, ada beberapa hal yang perlu diperbaiki. Data yang digunakan masih berasal dari satu lembaga bimbingan belajar saja, sehingga perlu diuji lagi apakah model ini bisa diterapkan secara luas pada populasi yang lebih besar mempertimbangkan. Penelitian ini juga hanya menggunakan tiga algoritma klasifikasi. Masih banyak algoritma lain yang bisa digunakan, seperti *XGBoost*, *Support Vector Machine*, atau *Deep Learning*, untuk membandingkan performa model. Penelitian berikutnya bisa menambahkan metode *explainable artificial intelligence* (XAI) seperti SHAP atau LIME. Dengan begitu, kita bisa lebih memahami faktor-faktor apa saja yang memengaruhi kelulusan siswa.

PEMBAHASAN

Penelitian ini secara keseluruhan berhasil membuktikan bahwa pendekatan *machine learning* berbasis CRISP-DM dapat diterapkan secara efektif untuk memprediksi kelulusan siswa SMA ke PTN dengan tingkat akurasi yang sangat tinggi. Temuan pokok penelitian ini menunjukkan bahwa algoritma *Random Forest* mencapai akurasi tertinggi sebesar 96,63%, diikuti *Decision Tree* (95,48%) dan *Naive Bayes* (94,06%). Peringkat kelas, nilai *try out*, akreditasi sekolah, dan kehadiran terbukti sebagai prediktor terkuat dalam menentukan peluang kelulusan siswa. Relevansi temuan ini sangat signifikan bagi praktisi pendidikan, khususnya lembaga bimbingan belajar, karena memberikan bukti empiris bahwa pembinaan siswa berbasis data bukan lagi sekadar wacana, melainkan sebuah pendekatan yang sudah matang untuk diimplementasikan. Kebaruan (*novelty*) utama penelitian ini terletak pada tiga aspek yang saling melengkapi. *Pertama*, penelitian ini mengintegrasikan variabel akademik, non-akademik, dan sosial-ekonomi secara komprehensif dalam satu model prediksi kelulusan PTN sebuah pendekatan holistik yang masih jarang dilakukan dalam konteks Indonesia. *Kedua*, penggunaan kerangka CRISP-DM yang menunjukkan bahwa secara terstruktur dan sistematis dalam konteks lembaga bimbingan belajar pra-PTN memberikan pendekatan metodologis yang lebih *rigorous* dibandingkan penelitian sebelumnya yang umumnya menggunakan pendekatan

ad-hoc. Ketiga, dataset yang digunakan mencakup periode pasca-pandemi COVID-19 (2021–2025), di mana terjadi perubahan signifikan sistem seleksi PTN dari SBMPTN menjadi SNBT dan SNBP, sehingga variabel yang teridentifikasi sebagai prediktor kuat mencerminkan kondisi seleksi yang lebih terkini dan relevan bagi para pemangku kepentingan pendidikan saat ini.

Hasil penelitian ini memiliki keterkaitan yang erat dengan berbagai temuan penelitian terdahulu di bidang *educational data mining* dan prediksi kinerja akademik. Penelitian Airlangga (2024) yang menggunakan model *deep learning* seperti MLP, CNN, dan LSTM untuk memprediksi kinerja akademik siswa secara umum menegaskan bahwa pendekatan berbasis data mampu menghasilkan prediksi yang andal sebuah landasan yang diperkuat oleh penelitian ini dalam konteks yang lebih spesifik. Sejalan dengan itu, AlOtaibi (2025) mendemonstrasikan bahwa *machine learning* dapat diintegrasikan ke dalam platform pendukung keputusan berbasis web untuk memantau kinerja siswa secara *real-time*, yang memperkuat potensi penerapan praktis model prediksi seperti yang dikembangkan dalam penelitian ini. Dari sisi pemilihan algoritma, Andrianof et al. (2025) menunjukkan bahwa *Random Forest* efektif dalam mengelola data akademik dengan pola hubungan yang kompleks di perguruan tinggi Indonesia. Temuan ini konsisten dengan hasil penelitian ini yang juga menempatkan *Random Forest* sebagai algoritma terbaik. Kaensar & Wongnri (2023) dalam konteks mahasiswa baru di Thailand juga menemukan keunggulan serupa, di mana *Random Forest* dengan *hyperparameter tuning* berhasil mengidentifikasi atribut penting yang memengaruhi keberhasilan akademik. Senada dengan itu, Al-Tameemi et al. (2024) menunjukkan bahwa *Random Forest* dan *Decision Tree* merupakan metode yang efektif dalam memprediksi performa akademik mahasiswa pada berbagai dataset pendidikan, dan mendukung temuan penelitian ini bahwa model berbasis *ensemble* seperti *Random Forest* mampu memberikan performa yang lebih baik dibandingkan model *Decision Tree* tunggal dalam menghasilkan prediksi yang akurat. Dari perspektif kajian sistematis, Issah et al. (2023) mengonfirmasi bahwa variabel pendidikan merupakan atribut paling konsisten dalam memengaruhi kinerja akademik di berbagai penelitian *machine learning*, yang mendukung dominannya faktor peringkat kelas dan nilai *try out* dalam penelitian ini. Guanin-Fajardo et al. (2024) menemukan bahwa faktor akademik memiliki kontribusi terbesar, konsisten dengan temuan penelitian ini. Nachouki & Abou Naaj (2022) juga menyimpulkan bahwa variabel akademik, riwayat nilai, serta capaian pembelajaran sebelumnya merupakan prediktor utama dalam memprediksi keberhasilan akademik mahasiswa, yang sejalan dengan hasil penelitian ini yang menunjukkan bahwa variabel peringkat kelas dan kehadiran memiliki kontribusi dominan dalam proses klasifikasi performa siswa. Lebih lanjut, Thakur & Rawat (2025) menunjukkan bahwa faktor perilaku belajar, riwayat akademik, dan karakteristik siswa menjadi atribut penting dalam membangun model prediksi performa akademik, sehingga memperkuat temuan penelitian ini bahwa kombinasi faktor akademik dan kehadiran siswa berperan signifikan dalam menentukan tingkat keberhasilan model. Arifin et al. (2026) turut menunjukkan bahwa menggabungkan faktor akademik, sosial-ekonomi, dan perilaku siswa dapat memperbaiki kualitas model klasifikasi kinerja akademik, meskipun dalam konteks yang berbeda dari penelitian ini. Terkait pengembangan model dan konteks pembelajaran yang beragam, de Oliveira et al. (2024) mengembangkan pendekatan *machine learning* untuk manajemen penerimaan mahasiswa di Jerman dan menemukan bahwa model prediktif dapat meningkatkan efisiensi proses seleksi. Okoye et al. (2024) mengembangkan model *ensemble* untuk prediksi retensi dan kelulusan mahasiswa, yang secara metodologis memiliki kesamaan dengan pendekatan *Random Forest* yang digunakan dalam penelitian ini. Wang & Yu (2025) dalam penelitiannya tentang prediksi kinerja mahasiswa dalam *online learning* turut memperkuat relevansi penggunaan *machine learning* dalam berbagai konteks pembelajaran. Ahmed et al. (2025) menggunakan *explainable machine learning* untuk prediksi performa akademik dan menemukan bahwa transparansi model sangat penting untuk adopsi di lingkungan pendidikan, yang memberikan perspektif berharga untuk pengembangan lebih lanjut dari penelitian ini.

Berdasarkan analisis komparatif, terdapat kesamaan dan perbedaan yang perlu diuraikan untuk memahami posisi penelitian ini dalam lanskap literatur yang lebih luas. Dari

sisi kesamaan, konsistensi keunggulan *Random Forest* atas algoritma lainnya yang ditemukan di hampir seluruh penelitian sejenis termasuk Andrianof et al. (2025); Kaensar & Wongnin (2023); Okoye et al. (2024) yang dapat dijelaskan oleh mekanisme *ensemble learning* yang menjadi fondasi algoritma ini. Kombinasi banyak pohon keputusan yang dibangun secara independen melalui proses *bagging* secara efektif mereduksi *variance* dan *bias* sekaligus, menghasilkan prediksi yang lebih *robust* terhadap *noise* dan heterogenitas dalam data pendidikan yang multivariat (Almalawi et al., 2024). Dominannya faktor akademik sebagai prediktor terkuat juga merupakan temuan yang konsisten lintas konteks, mencerminkan realitas universal bahwa kesiapan akademik merupakan modal utama dalam kompetisi seleksi perguruan tinggi di mana pun. Di sisi lain, terdapat beberapa perbedaan yang perlu dipahami secara kontekstual. *Pertama*, penelitian ini menemukan bahwa faktor sosial-ekonomi memiliki pengaruh yang relatif lebih rendah dibandingkan faktor akademik, berbeda dengan temuan Arifin et al. (2026) dan Ahmed et al. (2025) yang menemukan kontribusi sosial-ekonomi lebih substansial. Perbedaan ini kemungkinan besar disebabkan oleh karakteristik populasi: data penelitian ini berasal dari lembaga bimbingan belajar dengan profil siswa yang sudah memiliki motivasi akademik kuat dan secara aktif mempersiapkan diri, sehingga variasi sosial-ekonomi tidak sepengaruh pada populasi umum yang lebih heterogen. *Kedua*, akurasi *Random Forest* yang dicapai dalam penelitian ini (96,63%) relatif lebih tinggi dibandingkan beberapa studi sejenis seperti Azizah et al. (2024) dan Nimy et al., (2023), yang kemungkinan disebabkan oleh karakteristik data yang lebih homogen karena bersumber dari satu lembaga, serta kualitas data yang lebih terjaga melalui proses *preprocessing* yang komprehensif. *Ketiga*, penggunaan kerangka CRISP-DM dalam penelitian ini memberikan struktur metodologis yang lebih sistematis dibandingkan penelitian serupa yang menggunakan pendekatan *ad-hoc*, yang berkontribusi pada kualitas dan reproduktibilitas hasil penelitian.

Hasil penelitian ini juga memiliki beberapa kontribusi dan kebaruan yang signifikan terhadap bidang *educational data mining* dan *machine learning* di sektor pendidikan, khususnya dalam konteks Indonesia. Dari perspektif metodologis, penelitian ini menawarkan kerangka komprehensif berbasis CRISP-DM untuk pengembangan model prediksi kelulusan PTN yang mengintegrasikan tiga dimensi variabel secara bersamaan: akademik, non-akademik, dan sosial-ekonomi. Integrasi tiga dimensi ini merupakan pendekatan yang lebih holistik dibandingkan sebagian besar penelitian sebelumnya yang hanya fokus pada satu atau dua dimensi (Guanin-Fajardo et al., 2024; Issah et al., 2023). Dari perspektif substantif, penelitian ini memberikan bukti empiris bahwa peringkat kelas, akreditasi sekolah, nilai *try out*, dan kehadiran merupakan prediktor terkuat dalam menentukan kelulusan siswa ke PTN. Temuan ini memperbarui pemahaman tentang faktor-faktor penentu kelulusan PTN di Indonesia yang sebelumnya lebih banyak berfokus pada nilai akademik semata. Penelitian ini juga berkontribusi dalam memperkuat ekosistem *educational data mining* di Indonesia, yang masih relatif terbatas dibandingkan penelitian sejenis dari negara-negara maju. Selain itu, model yang dikembangkan memiliki potensi untuk diadaptasi dan diterapkan di berbagai lembaga bimbingan belajar lain di Indonesia, sehingga dapat memberikan dampak yang lebih luas dalam meningkatkan efektivitas pembinaan siswa menuju PTN. Lebih spesifik, kebaruan penelitian ini juga terletak pada penggunaan dataset yang bersumber dari periode pasca-pandemi COVID-19 (2021–2025), di mana terjadi perubahan signifikan dalam sistem seleksi PTN di Indonesia dari SBMPTN menjadi SNBT (Seleksi Nasional Berdasarkan Tes) dan SNBP (Seleksi Nasional Berdasarkan Prestasi). Konteks perubahan sistem seleksi ini memberikan dimensi tambahan yang membuat penelitian ini lebih relevan dan kontemporer dibandingkan penelitian serupa yang menggunakan data dari periode sebelumnya. Dengan demikian, variabel-variabel yang teridentifikasi sebagai prediktor kuat dalam penelitian ini mencerminkan kondisi seleksi PTN yang lebih terkini dan relevan bagi praktisi pendidikan saat ini. Penelitian ini juga membuktikan bahwa metodologi CRISP-DM yang awalnya dirancang untuk konteks industri dapat berhasil diadaptasi untuk konteks pendidikan di Indonesia, membuka peluang bagi penelitian metodologis lanjutan tentang adaptasi kerangka kerja *data mining* untuk sektor pendidikan. Hasil penelitian ini memiliki implikasi yang luas, baik dari perspektif praktis, akademis, maupun kebijakan. Dari perspektif praktis, model prediksi yang

dikembangkan dapat segera diimplementasikan oleh lembaga bimbingan belajar sebagai sistem pendukung keputusan berbasis data. Dengan model ini, pembimbing akademik dapat mengidentifikasi siswa yang membutuhkan perhatian khusus sejak awal masa bimbingan, merancang program pembinaan yang dipersonalisasi sesuai profil masing-masing siswa, serta mengalokasikan sumber daya bimbingan secara lebih efisien. Implikasi serupa juga berlaku bagi sekolah menengah atas, di mana model prediksi dapat digunakan sebagai alat bantu konseling akademik untuk memandu siswa dalam mempersiapkan diri menghadapi seleksi masuk PTN (W. Ahmed et al., 2025; AlOtaibi, 2025). Selanjutnya apabila dilihat dari perspektif kebijakan pendidikan, temuan penelitian ini dapat menjadi dasar bagi pemerintah dan pemangku kebijakan dalam merancang program intervensi yang lebih tepat sasaran. Hasil penelitian menunjukkan bahwa akreditasi sekolah memiliki pengaruh signifikan terhadap peluang kelulusan siswa ke PTN, yang menyiratkan pentingnya peningkatan mutu sekolah secara merata, terutama di daerah-daerah dengan proporsi sekolah berakreditasi rendah. Selain itu, temuan bahwa kehadiran berpengaruh positif terhadap kelulusan mengimplikasikan perlunya kebijakan yang mendorong keterlibatan aktif siswa dalam proses pembelajaran.

Temuan penelitian ini memiliki implikasi penting bagi desain kurikulum bimbingan belajar. Dari perspektif *praktis*, model prediksi yang dikembangkan dapat segera diimplementasikan oleh lembaga bimbingan belajar sebagai sistem pendukung keputusan berbasis data untuk mengidentifikasi siswa yang membutuhkan perhatian khusus sejak awal, merancang program pembinaan yang dipersonalisasi sesuai profil masing-masing siswa, serta mengalokasikan sumber daya bimbingan secara lebih efisien (W. Ahmed et al., 2025; AlOtaibi, 2025). Bagi sekolah menengah atas, model ini dapat difungsikan sebagai alat bantu konseling akademik yang berbasis data, bukan sekadar intuisi konselor. Dari perspektif *kebijakan*, temuan bahwa akreditasi sekolah berpengaruh signifikan terhadap peluang kelulusan siswa ke PTN menyiratkan urgensi peningkatan mutu sekolah secara merata, terutama di daerah dengan proporsi sekolah berakreditasi rendah. Temuan tentang peran kehadiran juga mengimplikasikan perlunya kebijakan yang mendorong keterlibatan aktif dan konsisten siswa dalam proses pembelajaran. Keduanya merupakan sinyal bagi pemerintah dan pemangku kebijakan untuk merancang intervensi yang lebih tepat sasaran guna meningkatkan ekuitas akses ke PTN. Dari perspektif akademis, penelitian ini membuka empat agenda penelitian lanjutan yang relevan: (1) penelitian berikutnya dapat memperluas cakupan data dengan melibatkan beberapa lembaga bimbingan belajar di berbagai kota, sehingga model yang dihasilkan lebih generalizable dan representative; (2) mengeksplorasi algoritma yang lebih canggih seperti XGBoost, *Support Vector Machine*, atau *Deep Learning* untuk perbandingan yang lebih komprehensif; (3) mengintegrasikan metode *explainable artificial intelligence* (XAI) seperti SHAP atau LIME guna meningkatkan transparansi model dan kemudahan adopsi oleh praktisi non-teknis (W. Ahmed et al., 2025; Jang et al., 2022); (4) menambahkan variabel psikologis seperti motivasi belajar, kecemasan akademik, dan *resilience* untuk memperkaya pemahaman tentang determinan keberhasilan siswa. Meskipun model yang dikembangkan dalam penelitian ini cukup akurat, penelitian ini juga memiliki keterbatasan yang perlu diakui: data berasal dari satu lembaga bimbingan belajar saja sehingga generalisasi ke populasi yang lebih luas perlu dilakukan dengan hati-hati, dan hanya tiga algoritma klasifikasi yang dibandingkan. Keterbatasan ini sekaligus menjadi peta jalan yang jelas bagi penelitian berikutnya untuk memperluas dan memperdalam kontribusi di bidang ini.

CONCLUSION

Penelitian ini berhasil mengembangkan model prediksi kelulusan siswa SMA ke Perguruan Tinggi Negeri (PTN) menggunakan pendekatan *machine learning* berbasis CRISP-DM dengan mengintegrasikan variabel akademik, non-akademik, dan sosial-ekonomi secara komprehensif. Hasil analisis menunjukkan bahwa algoritma *Random Forest* mencapai performa terbaik dengan rata-rata akurasi (96,63%), diikuti *Decision Tree* (95,48%) dan *Naive Bayes* (94,06%). Konsistensi kinerja *Random Forest* di berbagai skenario pembagian data (70:30, 80:20, 90:10) mengkonfirmasi keunggulan pendekatan *ensemble learning* dalam menangkap pola kompleks multivariat dalam data pendidikan. Variabel peringkat kelas, nilai *try out*, akreditasi

sekolah, dan kehadiran terbukti sebagai prediktor terkuat yang menentukan peluang kelulusan siswa ke PTN. Temuan ini menegaskan bahwa keberhasilan akademik siswa tidak hanya ditentukan oleh nilai akademik semata, tetapi juga oleh konsistensi belajar, kedisiplinan, dan kualitas lingkungan pendidikan.

Secara keseluruhan, penelitian ini memberikan kontribusi signifikan bagi pengembangan *educational data mining* di Indonesia dan mendukung adopsi sistem pendukung keputusan berbasis data di lembaga pendidikan. Model yang dikembangkan dapat digunakan sebagai dasar untuk menyusun strategi pembinaan yang lebih tepat sasaran, efisien, dan personal. Kontribusi utama penelitian ini meliputi: (1) pengembangan model prediksi komprehensif berbasis CRISP-DM yang mengintegrasikan tiga dimensi variabel; (2) bukti empiris tentang dominannya faktor akademik dan kedisiplinan dalam menentukan kelulusan PTN; (3) demonstrasi keunggulan algoritma *Random Forest* dalam konteks data pendidikan Indonesia; dan (4) kerangka metodologis yang dapat diadopsi oleh lembaga pendidikan lain. Penelitian selanjutnya disarankan untuk memperluas cakupan data ke beberapa lembaga, mengeksplorasi algoritma *machine learning* yang lebih beragam seperti XGBoost dan *deep learning*, serta mengintegrasikan metode *explainable AI* untuk meningkatkan transparansi serta penerimaan model prediksi oleh para praktisi pendidikan yang memiliki keterbatasan latar belakang teknis.

REFERENCES

- Abuzinadah, N., Umer, M., Ishaq, A., Al Hejaili, A., Alsubai, S., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2023). Role of convolutional features and machine learning for predicting student academic performance from MOODLE data. *PLOS ONE*, 18(11), e0293061. <https://doi.org/10.1371/journal.pone.0293061>
- Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, 10(1), 61–75. <https://doi.org/10.1108/JARHE-09-2017-0113>
- Ahmed, E. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Applied Computational Intelligence and Soft Computing*, 2024(1), 1–15. <https://doi.org/10.1155/2024/4067721>
- Ahmed, W., Wani, M. A., Plawiak, P., Meshoul, S., Mahmoud, A., & Hammad, M. (2025). Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Scientific Reports*, 15(1), 26879. <https://doi.org/10.1038/s41598-025-12353-4>
- Airlangga, G. (2024). Predicting Student Performance Using Deep Learning Models: A Comparative Study of MLP, CNN, BiLSTM, and LSTM with Attention. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(4), 1561–1567. <https://doi.org/10.57152/malcom.v4i4.1668>
- Al-Tameemi, G., Xue, J., Ali, I. H., & Ajit, S. (2024). A Hybrid Machine Learning Approach for Predicting Student Performance Using Multi-class Educational Datasets. *Procedia Computer Science*, 238, 888–895. <https://doi.org/10.1016/j.procs.2024.06.108>
- Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive Models for Educational Purposes: A Systematic Review. *Big Data and Cognitive Computing*, 8(12), 187. <https://doi.org/10.3390/bdcc8120187>
- AlOtaibi, N. F. (2025). A Web-based Decision Support Platform for Student Performance Prediction using Machine Learning. *International Journal of Advances in Artificial Intelligence and Machine Learning*, 2(3), 170–177. <https://doi.org/10.58723/ijaauml.v2i3.542>
- Andrianof, H., Aggy Pramana Gusman, & Okta Andrica Putra. (2025). Implementasi Algoritma Random Forest untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik: Studi Kasus di Perguruan Tinggi Indonesia. *Jurnal Sains Informatika Terapan*, 4(1), 24–28. <https://doi.org/10.62357/jsit.v4i2.464>
- Arifin, M., Nugraha, F., & Fithri, D. L. (2026). Student Performance Classification Using Academic, Socioeconomic, and Digital Behavior Features: A Comparative Study. *Journal of Information Systems and Informatics*, 8(1), 223–239. <https://doi.org/10.63158/journalisi.v8i1.1460>

- Azizah, Z., Ohyama, T., Zhao, X., Ohkawa, Y., & Mitsuishi, T. (2024). Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data. *Computers and Education: Artificial Intelligence*, 7, 100261. <https://doi.org/10.1016/j.caeai.2024.100261>
- Berka, P., & Marek, L. (2021). Bachelor's degree student dropouts: Who tend to stay and who tend to leave? *Studies in Educational Evaluation*, 70, 100999. <https://doi.org/10.1016/j.stueduc.2021.100999>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Caday, M. B., Dalaorao, G. A., & Abalorio, C. C. (2025). Academic Performance Analysis of Diploma Students Using Predictive Analytics. *Procedia Computer Science*, 257, 913–920. <https://doi.org/10.1016/j.procs.2025.03.117>
- Chen, J., Zhou, X., Yao, J., & Tang, S.-K. (2025). Application of machine learning in higher education to predict students' performance, learning engagement and self-efficacy: a systematic literature review. *Asian Education and Development Studies*, 14(2), 205–240. <https://doi.org/10.1108/AEDS-08-2024-0166>
- Cheng, B., Liu, Y., & Jia, Y. (2024). Evaluation of students' performance during the academic period using the XG-Boost Classifier-Enhanced AEO hybrid model. *Expert Systems with Applications*, 238, 122136. <https://doi.org/10.1016/j.eswa.2023.122136>
- Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T., & Gulam, S. B. (2025). Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors. *Systems and Soft Computing*, 7, 200352. <https://doi.org/10.1016/j.sasc.2025.200352>
- Delogu, M., Lagravinese, R., Paolini, D., & Resce, G. (2024). Predicting dropout from higher education: Evidence from Italy. *Economic Modelling*, 130, 106583. <https://doi.org/10.1016/j.econmod.2023.106583>
- de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2024). Interpretable Success Prediction in a Computer Networks Curricular Unit Using Machine Learning. *Procedia Computer Science*, 239, 598–605. <https://doi.org/10.1016/j.procs.2024.06.212>
- Diponegoro, M. H., Kusumawardani, S. S., & Hidayah, I. (2021). Tinjauan Pustaka Sistematis: Implementasi Metode Deep Learning pada Prediksi Kinerja Murid. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, 10(2), 131–138. <https://doi.org/10.22146/jnteti.v10i2.1417>
- Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, 5, 15991–16005. <https://doi.org/10.1109/ACCESS.2017.2654247>
- Guanin-Fajardo, J. H., Guña-Moya, J., & Casillas, J. (2024). Predicting Academic Success of College Students Using Machine Learning Techniques. *Data*, 9(4), 60. <https://doi.org/10.3390/data9040060>
- Hussain, S., & Khan, M. Q. (2023). Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning. *Annals of Data Science*, 10(3), 637–655. <https://doi.org/10.1007/s40745-021-00341-0>
- Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal*, 7, 100204. <https://doi.org/10.1016/j.dajour.2023.100204>
- Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies*, 27(9), 12855–12889. <https://doi.org/10.1007/s10639-022-11120-6>
- Kaensar, C., & Wongnin, W. (2023). Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(12), em2369. <https://doi.org/10.29333/ejmste/13863>
- Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4), 331–344. <https://doi.org/10.1007/s10462-011-9234-x>

- Maniyan, S., Ghousi, R., & Haeri, A. (2024). Data mining-based decision support system for educational decision makers: Extracting rules to enhance academic efficiency. *Computers and Education: Artificial Intelligence*, 6, 100242. <https://doi.org/10.1016/j.caeai.2024.100242>
- Matzavela, V., & Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. *Computers and Education: Artificial Intelligence*, 2, 100035. <https://doi.org/10.1016/j.caeai.2021.100035>
- Nachouki, M., & Abou Naaj, M. (2022). Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm. *International Journal of Distance Education Technologies*, 20(1), 1-17. <https://doi.org/10.4018/IJDET.296702>
- Nakhipova, V., Kerimbekov, Y., Umarova, Z., Bulbul, H. ibrahim, Suleimenova, L., & Adylbekova, E. (2024). Integration of Collaborative Filtering Into Naive Bayes Method to Enhance Student Performance Prediction. *International Journal of Information and Communication Technology Education*, 20(1), 1-18. <https://doi.org/10.4018/IJICTE.352512>
- Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying At-Risk Students for Early Intervention – A Probabilistic Machine Learning Approach. *Applied Sciences*, 13(6), 3869. <https://doi.org/10.3390/app13063869>
- Okoye, K., Nganji, J. T., Escamilla, J., & Hosseini, S. (2024). Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education. *Computers and Education: Artificial Intelligence*, 6, 100205. <https://doi.org/10.1016/j.caeai.2024.100205>
- Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 80, 3782-3785. <https://doi.org/10.1016/j.matpr.2021.07.382>
- Pecuchova, J., & Drlik, M. (2023). Predicting Students at Risk of Early Dropping Out from Course Using Ensemble Classification Methods. *Procedia Computer Science*, 225, 3223-3232. <https://doi.org/10.1016/j.procs.2023.10.316>
- Ravi, K., Kotecha, K., & Pawar, A. (2026). Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches. *Computers and Education: Artificial Intelligence*, 9, 100093. <https://doi.org/10.1016/j.caeai.2026.100093>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3). <https://doi.org/10.1002/widm.1355>
- Saputri, R. A., & Rosnita, L. (2025). A random forest-based predictive model for student academic performance: A case study in Indonesian public high schools. *Journal of Applied Informatics and Computing*, 9(1), 1-12. <https://doi.org/10.30871/jaic.v9i1.9460>
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, 526-534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Solano, J. A., Lancheros Cuesta, D. J., Umaña Ibáñez, S. F., & Coronado-Hernández, J. R. (2022). Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test. *Procedia Computer Science*, 198, 512-517. <https://doi.org/10.1016/j.procs.2021.12.278>
- Sultana, S., Khan, S., & Abbas, M. A. (2017). Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering & Education*, 54(2), 105-118. <https://doi.org/10.1177/0020720916688484>
- Sun, X., Pelet, J.-É., Dai, S., & Ma, Y. (2023). The Effects of Trust, Perceived Risk, Innovativeness, and Deal Proneness on Consumers' Purchasing Behavior in the Livestreaming Social Commerce Context. *Sustainability*, 15(23), 16320. <https://doi.org/10.3390/su152316320>
- Tan, C. J., & Tamayo, A. R. (2025). Academic performance analysis of diploma students using predictive analytics. *Procedia Computer Science*, 258, 1625-1636. <https://doi.org/10.1016/j.procs.2025.02.854>

- Thakur, S. N., & Rawat, B. (2025). Prediction of Student Performance in E-Learning Environment using Machine Learning Techniques. *Procedia Computer Science*, 259, 1416–1425. <https://doi.org/10.1016/j.procs.2025.04.096>
- Tran, T., Nguyen, G., Nguyen, D., & Le, A. (2023). Decision tree and random forest algorithms for student performance prediction: A comparative study. *Journal of Computer Science*, 19(2), 105–118. <https://doi.org/10.3844/jcssp.2023.105.118>
- Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104(1), 106189. <https://doi.org/10.1016/j.chb.2019.106189>
- Wang, J., & Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. *PLOS ONE*, 20(1), e0299018. <https://doi.org/10.1371/journal.pone.0299018>
- Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, 98, 137–144. <https://doi.org/10.1016/j.chb.2019.04.015>
- Yao, H., Tian, L., & Guo, L. (2023). Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Social Sciences & Humanities Open*, 8(1), 100521. <https://doi.org/10.1016/j.ssaho.2023.100521>
- Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89(1), 106903. <https://doi.org/10.1016/j.compeleceng.2020.106903>