

## Analitik Pendidikan 4.0: Penerapan Data Mining dalam Mengungkap Karakteristik Siswa

Tommy Jonathan Sinaga

Program Studi Teknik Informatika, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

E-Mail: [tommyjonathansinaga@outlook.com](mailto:tommyjonathansinaga@outlook.com)

### Article Info

#### Article history:

Received Jun 17, 2025

Revised Jul 20, 2025

Accepted Aug 10, 2025

#### Kata Kunci:

Analitik Pendidikan 4.0

Data Mining

Karakteristik Siswa

Pohon Keputusan

Klasifikasi

#### Keywords:

Education Analytics 4.0

Data Mining

Student Characteristics

Decision Trees

Classification

### ABSTRAK

Salah satu tantangan paling menarik dan kompleks dalam Penambangan Data Pendidikan (EDM) adalah menganalisis kinerja siswa. Peneliti sangat tertarik pada bidang ini karena berbagai faktor yang memengaruhi hasil belajar, serta banyaknya data yang tersedia, terutama dalam konteks pembelajaran yang ditingkatkan teknologi. Meskipun banyak penelitian ada dalam EDM, hanya sedikit yang berfokus secara khusus pada evaluasi dan prediksi prestasi siswa. Sebagian besar survei bertujuan untuk mengidentifikasi pola atau faktor yang dapat memprediksi kinerja siswa. Penelitian ini mengusulkan penggunaan algoritma penambangan data untuk mengekstrak data yang relevan dan akurat untuk analisis lebih lanjut. Dalam tinjauan pustaka ini, penulis mengkaji pendekatan yang ada di bidang ini dan menyoroti bagaimana interaksi siswa dengan sistem manajemen pembelajaran dan data tugas penilaian dapat memberikan wawasan berharga untuk prediksi kinerja awal. Penulis juga mengidentifikasi peran penting yang dimainkan oleh jenis sistem pendidikan dalam membentuk proyeksi awal prestasi siswa.

### ABSTRACT

One of the most exciting and complex challenges in Educational Data Mining (EDM) is analyzing student performance. Researchers are very interested in this field because of the various factors that influence learning outcomes, as well as the abundance of available data, especially in the context of technology-enhanced learning. While much research exists in EDM, few focus specifically on the evaluation and prediction of student achievement. Most surveys aim to identify patterns or factors that can predict student performance. This study proposes the use of data mining algorithms to extract relevant and accurate data for further analysis. In this literature review, author examine existing approaches in this field and highlight how student interactions with learning management systems and assessment task data can provide valuable insights for early performance prediction. The authors also identify the important role that the type of education system plays in shaping initial projections of student achievement.

This is an open access article under the [CC BY-NC](https://creativecommons.org/licenses/by-nc/4.0/) license.



#### Corresponding Author:

Tommy Jonathan Sinaga,

Informatika, STIKOM Tunas Bangsa

Jl. Jendral Sudirman Blok A No.1,2,3 Pematangsiantar

Email: [tommyjonathansinaga@outlook.com](mailto:tommyjonathansinaga@outlook.com)

## 1. PENDAHULUAN

Perkembangan Teknologi Informasi dan Komunikasi (TIK) telah membawa perubahan besar dalam dunia pendidikan, khususnya melalui penerapan *online learning* (OL). Model pembelajaran ini mampu meningkatkan pengalaman belajar siswa sekaligus mengurangi ketergantungan pada interaksi langsung dengan pengajar. Aktivitas dalam OL, seperti akses materi, diskusi daring, kuis, serta pencatatan aktivitas pengguna, menghasilkan data dalam jumlah besar dan beragam yang dapat dimanfaatkan untuk analisis lebih lanjut

(Marlina et al., 2024). Untuk mengolah data tersebut, diperlukan teknik *data mining*, yaitu proses untuk menemukan pola atau informasi tersembunyi dari kumpulan data yang besar. Teknik ini tidak hanya digunakan dalam bidang bisnis, tetapi juga memberikan kontribusi penting dalam pengembangan ilmu pengetahuan, termasuk dalam bidang pendidikan (Huang et al., 2021; Nurhayati & Lawanda, 2023).

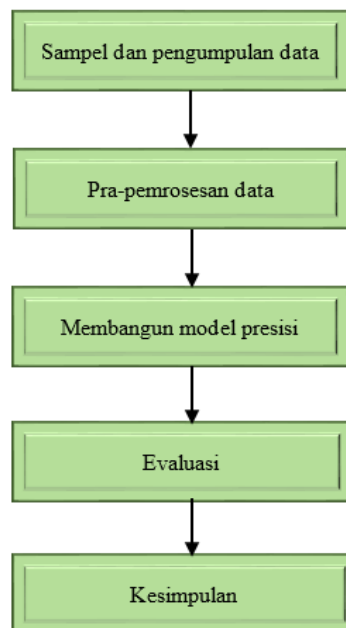
Dalam bidang pendidikan, analisis data mahasiswa menjadi hal yang penting, terutama untuk memahami perilaku dan memprediksi prestasi akademik. Dengan adanya prediksi ini, institusi pendidikan dapat melakukan pemantauan secara lebih terstruktur serta memberikan penanganan lebih awal bagi mahasiswa yang berpotensi mengalami penurunan kinerja (Khan & Ghosh, 2018). Penelitian ini bertujuan untuk mengkaji penerapan teknik data mining dalam menganalisis serta memprediksi kinerja akademik mahasiswa menggunakan data historis. Permasalahan ini termasuk dalam kategori *supervised learning*, karena data yang digunakan telah memiliki label atau target yang jelas (Kabakchieva, 2013). Berbagai metode klasifikasi, seperti Decision Tree, Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), dan K-Nearest Neighbors (KNN), telah banyak digunakan dalam penelitian sebelumnya dengan tingkat akurasi yang berbeda-beda (Kumar et al., 2017).

Sejumlah penelitian menunjukkan bahwa teknik data mining mampu mengelompokkan mahasiswa berdasarkan tingkat prestasi akademik, mulai dari kategori sangat baik hingga rendah. Metode Decision Tree, misalnya, terbukti memiliki tingkat akurasi yang cukup tinggi dibandingkan metode lainnya (Asif et al., 2017). Selain itu, pendekatan berbasis *big data* juga mendukung berbagai kebutuhan analisis pendidikan, seperti identifikasi risiko akademik, pemberian rekomendasi pembelajaran, serta visualisasi data untuk membantu pengambilan keputusan (Buenaño-Fernández et al., 2019). Penerapan data mining di perguruan tinggi juga membantu dalam meningkatkan kualitas pengelolaan pendidikan dan layanan kepada mahasiswa. Analisis data akademik, termasuk pada bidang tertentu seperti System Application and Product (SAP), menjadi penting untuk mempersiapkan lulusan yang sesuai dengan kebutuhan dunia kerja (Aziz & Ahmad, 2014).

Penelitian ini difokuskan pada analisis kinerja mahasiswa melalui dua pendekatan utama, yaitu evaluasi hasil akademik dan identifikasi perkembangan mahasiswa selama masa studi. Hasil penelitian diharapkan dapat membantu pendidik dalam memberikan perhatian yang lebih tepat kepada mahasiswa, baik yang berprestasi tinggi maupun yang membutuhkan dukungan akademik, sehingga proses pembelajaran dapat berjalan lebih efektif (Yağcı, 2022).

## 2. METODE PENELITIAN

Berikut ini metode penelitian yang digunakan dapat dilihat pada Gambar 1 yang terdiri dari pengumpulan sampel dan data, pra-pemrosesan data, membangun model presisi, evaluasi, dan tahap akhir adalah kesimpulan:



Gambar 1. Metode Penelitian

## 2.1 Sampel dan Pengumpulan Data

Penelitian yang dilakukan oleh universitas vokasi dan sains di Indonesia pada akhir tahun ajaran perdana 2020–2021. Prosedur penilaian dan pendaftaran sekolah digunakan untuk mengumpulkan data untuk *prototipe*. Siswa harus menggunakan teknologi pendaftaran sekolah untuk mengisi aplikasi terkomputerisasi dengan data pribadi mereka ketika pertama kali diterima di institusi. Kemudian, seiring setiap siswa maju melalui studi mereka, nilai dan prestasi mereka dimasukkan ke dalam sistem evaluasi sekolah. Selain itu, profil yang dipertahankan dari setiap siswa pada awal penelitian menggabungkan delapan belas variabel yang berkaitan dengan keadaan pribadi mereka dan satu parameter yang mencerminkan nilai rata-rata dari semua peringkat bidang yang berbeda, yang telah mereka capai di semester awal.

## 2.2 Pra-pemrosesan data

Penulis melakukan pembersihan data dan teknik harmonisasi data pada tahap pra-pemrosesan data. Setelah 18 variabel yang membentuk output (kinerja belajar) ditentukan, kami mengurutkan data kategori dan memperbaiki kasus nilai yang hilang dalam tahap pembersihan data. Tahap ini mencakup pengkodean data grup dan penolakan contoh dengan nilai yang hilang. Persamaan (1) digunakan untuk menormalkan data pada tahap pemurnian data.

$$Y_{mon} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \quad (1)$$

Dengan  $Y_{min}$ ,  $Y_{max}$ , dan  $Y_{mon}$  menunjukkan nilai minimum, maksimum, dan yang dinormalisasi, secara berturut-turut.

## 2.3 Pra-pemrosesan data

Eksperimen yang termasuk dalam penelitian dilakukan menggunakan RAM 64 GB dan CPU *Intel (R) Xeon (R) E-2174G 3.80 GHz* yang menjalankan sistem operasi *Windows*. Tergantung pada MLP, RF, dan komputasi DT, empat model pembelajaran terkontrol dikembangkan. Python diimplementasikan untuk membuat model prediksi RF, sedangkan strategi C5.0 dan *CART* digunakan untuk menghasilkan model prediksi DT. Untuk semua model, investigasi dijalankan sebanyak lima kali. Nilai rata-rata dan standar *error* dari hasil pengurutan dalam setiap kasus kemudian digunakan sebagai dasar ketika menilai komputasi DT dan RF. Tujuan utama dari beberapa proyek penelitian adalah untuk menemukan atribut yang memiliki efek besar pada kinerja belajar siswa dan menganalisis serta membandingkan efektivitasnya dalam memprediksi keberhasilan akademis pendatang baru pada *dataset*.

Selain itu, investigasi saat ini menggabungkan tiga kemungkinan jenis informasi produk berikut:

- Kasus 1 adalah dasar untuk hasil, yaitu klasifikasi Sangat Baik, Baik Sekali, Baik, Rata-rata, dan Buruk yang digunakan untuk menghitung keandalan perkiraan awal dan keseluruhan dari empat model.
- Kasus 2 melibatkan penggabungan kelas Sangat Baik, Baik, dan Cukup dari sebagian besar keluaran menjadi kelas Standar untuk menentukan apakah dari keempat pendekatan tersebut ada yang dapat mempertimbangkan minoritas.
- Kasus 3 mempersempit sorotan ke pencapaian periode yang lebih kecil, Luar Biasa dan Buruk.

## 2.4 Pohon Keputusan

Langkah-langkah yang disebutkan di atas adalah inti dari proses pengujian algoritma C5.0 untuk masing-masing dari tiga situasi dalam penyelidikan ini. Siapkan variabel pohon keputusan, membuat pendahuluan *Decision Tree*, membuat informasi dari pengujian dan pelatihan, pertahankan pohon, lalu saring pohon yang dipangkas untuk membuatnya lebih mudah dipahami. Sekarang pilih pohon yang menghasilkan yang terbaik di antara semua pohon yang dibuat. Kemudian ulangi langkah 1–6 untuk sepuluh pengujian dan gunakan standar deviasi dan median dari akurasi klasifikasi dalam sepuluh percobaan sebagai tolok ukur. Dengan menggunakan penyelidikan validasi silang *10-fold*, kami menghasilkan DT menggunakan prosedur C5.0 untuk setiap lipatan dalam kumpulan data. Sepuluh set berukuran sama dibuat dari set informasi yang diambil, dan setiap set berfungsi sebagai set pengujian secara bergantian. Bersamaan dengan set pengujian, kami mengembangkan DT yang memanfaatkan sembilan set lainnya sebagai materi pelatihan kami. Akibatnya, kami memiliki sepuluh pohon. Setelah memilih pohon dengan efektivitas terbaik, semua karakteristik yang masih ada dinilai identik.

Investigasi saat ini menerapkan strategi *CART* dengan Python sebagai teknik lebih lanjut untuk menguji, mengukur, dan menginterpretasikan presisi prediksi dan pilihan fitur signifikan antara C5.0 dan *CART*, setelah mendapatkan hasil percobaan DT. Untuk setiap dari tiga situasi, sistem *CART* diverifikasi menggunakan teknik berikut: Hasilkan data pengujian dan pelatihan awalnya. Kemudian, perlu untuk mengkonfigurasi peraturan DT untuk memulai metode. Untuk prediksi yang akurat, analisis DT menggunakan validasi silang, pengujian, dan pelatihan. Memplot hasil dari ukuran kepentingan Gini adalah langkah selanjutnya. Sepanjang sepuluh

percobaan, ulangi langkah 1-4 untuk hasil yang lebih baik. Manfaatkan nilai standar deviasi dan rata-rata dari kinerja kategorisasi dari uji lapangan sebagai tolok ukur.

## 2.5 Model Data Mining

Analisis regresi, serta klasifikasi, adalah tujuan *data mining* yang vital dan dualistik. Saling melibatkan pembelajaran yang diawasi, di mana prediksi dimodifikasi agar sesuai dengan satu set informasi dari  $l \in \{1, \dots, O\}$  contoh masing-masing memplot vektor sebagai input  $(y_1^l, \dots, y_j^l)$  ke target yang ditentukan  $y_l$ . Dari persamaan di bawah ini (2), nilai-nilai yang diberikan disebutkan dengan jelas. Perbedaan utama berkaitan dengan penggambaran hasil, yang bersifat independen dalam kasus pengurutan dan tak berkesudahan dalam situasi regresi. *Root Mean Squared* (RMSE) adalah metrik yang umum digunakan dalam analisis regresi, sedangkan *Percentage of Correct Classifications* (PCC) sering digunakan dalam menilai model prediktif untuk kategorisasi. Klasifikasi yang baik ditunjukkan oleh PCC yang tinggi (yaitu, sangat mendekati 100%) (3), dan analisis regresi harus merekomendasikan kesalahan global yang rendah (yaitu, RMSE mendekati nol) menggunakan persamaan (4). Pengukuran ini dapat diperoleh dengan menggunakan persamaan:

$$\varphi(j) = \begin{cases} 1, & \text{if } z_j = \hat{z}_j \\ 0, & \text{else} \end{cases} \quad (2)$$

$$PCC = \sum_{j=1}^M \varphi(j) / O \times 100(\%) \quad (3)$$

$$RMSE = \sqrt{\sum_{j=1}^O (z_j - \hat{z}_j)^2 / O} \quad (4)$$

Di mana nilai yang diantisipasi untuk ilustrasi ke- $j$  ditunjukkan oleh  $\hat{z}_j$ . Tiga metode pengawasan diterapkan saat memodelkan nilai Matematika dalam fungsi ini:

1. Klasifikasi biner: Siswa dianggap lulus jika nilai akhirnya (G3) 10 atau lebih, dan dianggap gagal jika nilainya di bawah 10.
2. Pengelompokan 5-Level: Nilai akhir siswa juga dikelompokkan ke dalam 5 level, berdasarkan konversi nilai Erasmus (lihat Tabel 1).
3. Analisis regresi: Digunakan untuk memprediksi nilai akhir (G3), yang bisa berada dalam rentang 0 sampai 20.

Tabel 1. Sistem klasifikasi lima tingkat

	I	II	III	IV	V
Negara	(sangat bagus/sangat baik)	(Bagus)	(Memuaskan)	(Memadai)	(Gagal)
Prancis	17-21	15-16	13-14	11-12	1-10
Indonesia	B	C	D	E	G

Untuk tugas klasifikasi dan analisis regresi, beberapa jenis teknik *data mining* telah diajukan, masing-masing dengan fitur dan tujuan yang berbeda. Satu set pedoman dilambangkan oleh *decision tree* (DT), pengaturan terurut yang menggunakan tingkatan hierarki untuk membedakan antara nilai. Satu set aturan *If-then* yang intuitif dapat ditarik dari representasi visual ini. Metode *data mining* yang belum lengkap, sebuah pengaturan yang dikenal sebagai *Random Forest* (RF). Pilihan fitur acak dari bagian keseluruhan pelatihan *bootstrap* merupakan dasar untuk setiap pohon, dan evaluasi RF dibuat dengan merata-ratakan hasil *Decision Tree*. Jika dibandingkan dengan hanya satu *data mining*, sinyal frekuensi radio lebih sulit untuk dipahami, tetapi tetap memungkinkan untuk memberikan data deskriptif tentang pentingnya nilai *input*. Untuk kewajiban data *mining*, metode *nonlinier* seperti *Support Vector Machine* (SVM) dan *Neural Network* (NN) juga telah ditawarkan; fungsi-fungsi tersebut berkinerja lebih efisien ketika ada tingkat ketidakteraturan yang tinggi. *Perceptron multilayer* (MLP) yang banyak digunakan berfungsi sebagai tulang punggung untuk model NN dalam penelitian ini, yang memiliki lapisan tersembunyi dan *node* H tersembunyi. SVM menggunakan penyaring gaussian yang memiliki satu hiperparameter ( $\beta$ ). Perlu diakui bahwa NN dan SVM menggunakan ilustrasi model yang sulit dipahami oleh manusia. Selanjutnya, mengingat bahwa teknik *data mining*, *random forest* secara khusus mencapai pemilihan atribut internal, mereka kurang terpengaruh oleh bahan yang tidak signifikan dibandingkan dengan NN dan SVM (Gori, 2024).

### 3. HASIL DAN PEMBAHASAN

Pengklasifikasi yang paling andal digabungkan ke dalam program perangkat lunak yang sederhana untuk memprediksi prestasi siswa, sehingga lebih mudah bagi instruktur untuk mengenali siswa yang kesulitan dan menyarankan tindakan korektif.

#### 3.1 Dataset

Informasi yang dikumpulkan dalam set penelitian mencakup pencapaian siswa Universitas tahun pertama dalam matematika, khususnya mereka yang berusia antara 14 dan 15 tahun. Lyceum swasta "Avgoulea-Linardatou" mengumpulkan data, yang mencakup 279 struktur berbeda, antara tahun 2007 dan 2010. Fitur-fitur tersebut termasuk data mengenai pencapaian siswa, yang melibatkan nilai dari ujian lisan, dan penilaian akhir. Kumpulan fitur yang disediakan dalam Tabel 2 dapat dibagi menjadi dua kelompok utama berdasarkan kinerja siswa selama semester pertama dan kedua, sesuai. Selain itu, menurut sistem pengelompokan yang diterapkan pada evaluasi instruksional di sekolah-sekolah Yunani, siswa dipisahkan menjadi empat tingkatan:

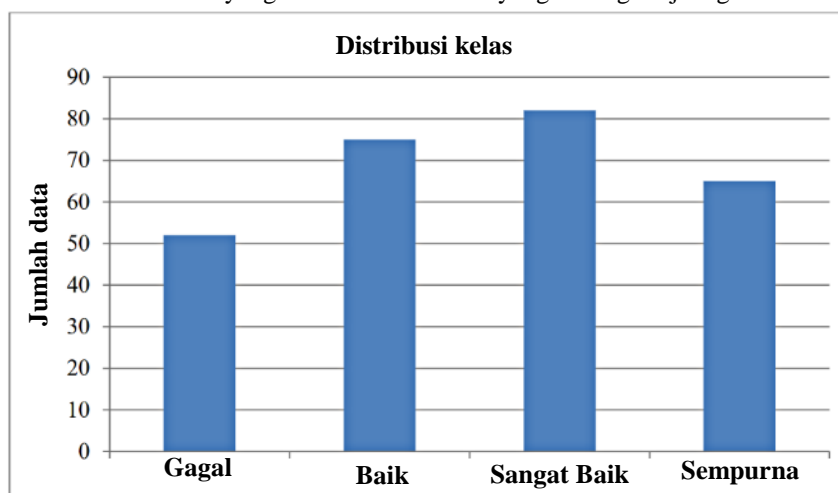
- Nilai "Gagal" mewakili skor siswa antara 0 dan 9.
- Nilai "Baik" menunjukkan pencapaian individu dalam hierarki 10 hingga 14.
- Pencapaian akademik siswa antara 15 dan 17 dianggap nilai "sangat baik."
- Setiap pencapaian siswa antara 18 dan 20 dapat dianggap nilai "sangat baik."

Tabel 2. Daftar parameter yang digunakan dalam studi penelitian

Atribut Mahasiswa ½ Semester	Rentang Nilai
Nilai lisan	[0,30]
Nilai Ujian 1	[0,30]
Nilai Ujian 2	[0,30]
Nilai ujian akhir	[0,30]
Nilai penutup	[0,30]

Penyebaran kelas terlihat pada Gambar 2, yang menampilkan bagian-bagian yang dikategorikan sebagai nilai "Gagal" (53 kasus), nilai "Baik" (76 kasus), nilai "Sangat baik" (85 kasus), dan nilai "Sempurna" (65 kasus). Dua set data telah dirancang dengan bantuan distribusi kelas dan ciri-ciri yang ditunjukkan dalam Tabel 1, karena informasi ini penting bagi seorang guru untuk mengetahui siswa yang kurang baik di tengah semester pendidikan.

- DATA<sub>A</sub>*: Ini menggabungkan kualitas yang memengaruhi pencapaian semester pertama siswa.
- DATA<sub>AB</sub>*: Termasuk atribut yang terkait dengan pencapaian semester pertama dan kedua siswa. Terlihat bahwa setiap kumpulan data dalam investigasi kami digunakan untuk membangun pengklasifikasi individual yang membedakan siswa yang sedang berjuang.



Gambar 2. Distribusi kelas

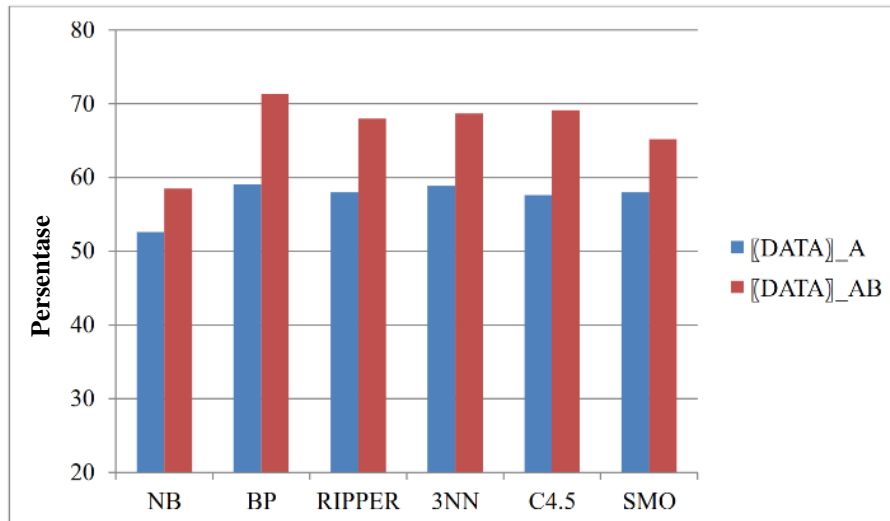
Penulis melakukan berbagai pengujian untuk mengetahui sistem pembelajaran mana yang paling baik mengantisipasi nilai siswa (yaitu, "Gagal," "Baik," "Sangat baik," atau "Sangat baik") menggunakan nilai mereka dari semester akademik dan ekstrakurikuler. Akibatnya, penulis memilih teknik yang paling banyak dinikmati dan sering diajukan untuk setiap prosedur *machine learning* yang telah dieksplorasi. Implementasi

klas algoritma Bayesian adalah pendekatan *Naive Bayes* (NB), yang paling sering diterapkan. Diberikan keadaan fitur kelas, ini adalah teknik pembelajaran langsung yang mengasumsikan bahwa setiap karakteristik tidak memengaruhi yang lain. Algoritma pembelajaran yang dibangun untuk menghasilkan jaringan saraf, algoritma *backpropagation* (BP) dengan momentum, adalah representasi yang efektif dari ANNs. Karena teknik Ripper adalah salah satu teknik yang paling umum digunakan untuk memberikan aturan klasifikasi, telah dipilih sebagai ilustrasi strategi aturan pembelajaran. Operasi perolehan data adalah intuisi pertumbuhan yang digunakan oleh Ripper, yang menghasilkan aturan dengan terus-menerus mengembangkan dan memotong. Sementara itu, kami menggunakan teknik *3-Nearest Neighbors* (3NN) sebagai audiens berbasis contoh, menggunakan Jarak Euclides sebagai sistem metrik pengukuran. Dengan algoritma C4.5 adalah yang lebih khas dalam tinjauan penulis yang diambil dari pohon keputusan. Algoritma C4.5 memilih elemen mana yang paling baik memecah contoh pelatihan pada setiap tahap proses pemisahan dengan menggunakan properti empiris yang disebut perolehan pengetahuan. Tabel 3 menawarkan ikhtisar pencapaian setiap pengklasifikasi berdasarkan jumlah urutan yang diklasifikasikan dengan benar dalam kumpulan data yang disediakan. Jelas bahwa tidak ada algoritma yang dapat terus-menerus melampaui yang lain dalam hal pencapaian. Lebih khusus lagi, BP (*backpropagation*) menunjukkan proporsi terbesar dari contoh yang diidentifikasi dengan benar tentang *dataset DATA<sub>A</sub>*, sementara 3NN menunjukkan hasil terbaik dalam hal *dataset DATA<sub>AB</sub>*.

Tabel 3. Akurasi klasifikasi pada setiap dataset

Klasifikasi	Dataset	
	(%) <i>DATA<sub>A</sub></i>	(%) <i>DATA<sub>AB</sub></i>
NB	52.6	58.5
BP	59.1	71.3
Ripper	58.0	68.0
3NN	58.9	68.7
C4.5	57.6	69.1
SMO	58.0	65.2

Akurasi pengklasifikasi juga direpresentasikan menggunakan grafik berikut pada Gambar 4. yang akan memberikan detail jelas mengenai akurasi pengklasifikasi.



Gambar 4. Histogram untuk akurasi Klasifikasi

Tujuan utama penulis adalah untuk mengembangkan klasifikasi kejadian yang lebih spesifik dan benar, jadi penulis menggunakan pemungutan suara, peraturan, dan metodologi penilaian untuk menggabungkan hasil yang diantisipasi dari berbagai prosedur pada *dataset* yang telah ditetapkan. Kalimat berikutnya menunjukkan fungsi dari strategi yang tercantum di kolom pertama Tabel 4:

- (BestCV) *Best Cross-Validation* adalah singkatan dari prosedur preferensi unggul pengklasifikasi.
- Voting* menunjukkan metode pemungutan suara yang mudah relatif yang mengintegrasikan estimasi dari berbagai teknik yang ditunjukkan dalam Tabel 3.
- Stacked Generalization* adalah singkatan dari prosedur *stacking*, yang menggunakan *Multiple Linear Regression* (MLR) sebagai algoritma meta-level dan pengklasifikasi dasar yang sama seperti pemilihan.

- d. *Grading* adalah bentuk singkat dari prosedur *grading* yang memanfaatkan pengklasifikasi utama situasi *10-Nearest Neighbors* sebagai pengklasifikasi meta-level dan pengklasifikasi dasar yang sama seperti *voting*.
- e. *Voting\** menyiratkan proses pemungutan suara yang relatif mudah yang terdiri dari SMO, BP, 3NN, dan Ripper sebagai pengklasifikasi dasar.
- f. Kata "*stacking*" sesuai dengan metodologi *stacking* yang melibatkan penggunaan MLR sebagai pengklasifikasi meta-level dan pengklasifikasi titik awal yang sama dengan *Voting*.
- g. *Grading\** mengacu pada prosedur evaluasi yang menggunakan 10-NN sebagai pengklasifikasi dasar contoh dan *Voting* sebagai kerangka kerja untuk algoritma meta-level.

Tabel 4. Presisi Akurasi untuk setiap *dataset*

Klasifikasi	Dataset	
	(%) <i>DATA<sub>A</sub></i>	(%) <i>DATA<sub>AB</sub></i>
<i>Best CV</i>	60.9	71.4
<i>Voting</i>	59.2	87.1
<i>Stacking</i>	57.7	69.6
<i>Grading</i>	58.8	72.8
<i>Voting*</i>	61.7	91.4
<i>Stacking*</i>	58.8	72.8

Berdasarkan pemahaman Tabel 4, pendekatan *Voting* melampaui kombinasi lainnya dalam hal akurasi, dengan *Voting Technique Method* mengonfirmasi efektivitas terbaik di seluruh bidang untuk kedua *dataset* (Livieris et al., 2016).

#### 4. KESIMPULAN

Memanfaatkan metode matematika pembelajaran mesin dan penambahan data untuk estimasi adalah instrumen efektif yang membantu instruktur dalam mengenali siswa yang cenderung berkinerja buruk sejak dini dan merupakan awal yang tak ternilai dalam strategi intervensi mereka. Kami membangun studi kasus tentang ujian matematika akhir tahun pertama Universitas, serta alat pendukung keputusan yang mudah digunakan untuk meramalkan kinerja siswa dalam proyek ini. Untuk memprediksi efektivitas belajar siswa, kami menggunakan elemen latar belakang keluarga dalam penelitian ini, yang dapat diperoleh pada awal musim mahasiswa baru. Segera setelah pendatang baru memulai sekolah, kami dapat menggunakan model yang ada untuk memperkirakan kinerja akademis mereka. Dengan bantuan pendekatan *Data Mining* ini, sangat mudah untuk memeriksa kelas yang diperoleh pelajar untuk kegiatan akademis maupun ekstrakurikuler. *Decision Tree* (DT), *Random Forest* (RF), *Neural Network* (NN), dan *Support Vector Machine* (SVM) adalah empat metode DM yang dievaluasi. Tiga tujuan *Data Mining* independen (klasifikasi 5 tingkat dan regresi) juga diperiksa. Berbagai opsi masukan (seperti menggabungkan atau menghapus nilai sebelumnya) juga dieksplorasi. Temuan yang dikumpulkan menunjukkan bahwa jika hasil untuk kerangka waktu sekolah pertama dan atau kedua dapat diakses, prediksi yang sangat akurat dapat dicapai. Ini membantu temuan penelitian bahwa pertunjukan sebelumnya memiliki panduan penting tentang pencapaian bagi siswa. Untuk meningkatkan basis data siswa, penulis juga memperluas pengujian penelitian ke institusi pendidikan dan tahun kelas lainnya. Sebagai konsekuensi dari penggunaan teknik *data mining* ini, penyelidikan data pendidikan siswa menjadi tepat dan jelas.

#### REFERENCES

- Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Comput. Educ.*, *113*, 177–194. <https://api.semanticscholar.org/CorpusID:26870758>
- Aziz, A. B. A., & Ahmad, N. (2014). *First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms*. <https://api.semanticscholar.org/CorpusID:111383032>
- Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability*. <https://api.semanticscholar.org/CorpusID:181333905>
- Gori, T. (2024). Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa. *Jurnal Teknologi Informasi Dan Ilmu Komputer*. <https://api.semanticscholar.org/CorpusID:268197786>

- Gunasekara, S., & Saarela, M. (2024). Explainability in Educational Data Mining and Learning Analytics: An Umbrella Review. *Educational Data Mining*. <https://api.semanticscholar.org/CorpusID:273692026>
- Huang, Y., Liu, H., & Pan, J. (2021). Identification of data mining research frontier based on conference papers. *Int. J. Crowd Sci.*, 5, 143–153. <https://api.semanticscholar.org/CorpusID:236284849>
- Huynh-Cam, T.-T., Chen, L.-S., & Le, H. (2021). Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. *Algorithms*, 14, 318. <https://api.semanticscholar.org/CorpusID:240473868>
- Kabakchieva, D. (2013). *Predicting Student Performance by Using Data Mining Methods for Classification*. <https://api.semanticscholar.org/CorpusID:7641809>
- Khan, A., & Ghosh, S. K. (2018). Data mining based analysis to explore the effect of teaching on student performance. *Education and Information Technologies*, 23, 1677–1697. <https://api.semanticscholar.org/CorpusID:7919350>
- Kumar, M., Singh, Prof. A. J., & Handa, D. (2017). Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques. *International Journal of Education and Management Engineering*, 7, 40–49. <https://api.semanticscholar.org/CorpusID:67421239>
- Livieris, I. E., Mikropoulos, T., & Pintelas, P. E. (2016). *A decision support system for predicting students' performance*. <https://api.semanticscholar.org/CorpusID:4092776>
- Marlina, R., Zaharuddin, Ngemba, H. R., Smith, J., Perangkat, P., Efisiensi, L., Keamanan, O., & Infrastruktur, D. O. (2024). Manfaat Integrasi IoT dalam Pengembangan Perangkat Lunak di Sektor Pendidikan. *Jurnal MENTARI: Manajemen, Pendidikan Dan Teknologi Informasi*. <https://api.semanticscholar.org/CorpusID:277158450>
- Nurhayati, E. S., & Lawanda, I. I. (2023). Perkembangan dan Tren Penelitian Global tentang Research Data Management. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi Dan Kearsipan*. <https://api.semanticscholar.org/CorpusID:267129157>
- Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9. <https://api.semanticscholar.org/CorpusID:247233325>