

Penerapan Algoritma DBSCAN dan XGBoost untuk Menganalisis Keberhasilan Pembelajaran Bahasa Indonesia bagi Penutur Asing (BIPA)

Nadya Rahma¹, Halim Maulana²

^{1,2}Universitas Muhammadiyah Sumatera Utara, Indonesia

*nadyarahma2233@gmail.com

Submit
10 Februari 2026

Review
13 Februari 2026

Publish
11 Maret 2026

Abstrak

Evaluasi tradisional keberhasilan pembelajaran Bahasa Indonesia bagi Penutur Asing (BIPA) masih bergantung pada penilaian subjektif pengajar yang kurang objektif. Penelitian ini bertujuan mengintegrasikan algoritma DBSCAN dan XGBoost untuk menganalisis pola pembelajaran dan mengidentifikasi faktor dominan keberhasilan BIPA. Pendekatan Educational Data Mining (EDM) kuantitatif eksploratif diterapkan pada 200 siswa Songsermasana School, Hat Yai, Thailand, selama program KKN 27 hari. Instrumen mencakup skor aktivitas, partisipasi kelas, kehadiran, nilai ujian, nilai kuis, waktu belajar, dan frekuensi pertanyaan; teknik analisis meliputi preprocessing, klusterisasi DBSCAN ($\text{eps}=0,5$, $\text{minPts}=5$), dan analisis feature importance XGBoost ($\text{learning rate}=0,1$, $\text{n_estimators}=100$, $\text{max_depth}=5$). Hasil menunjukkan tiga kluster: Cluster 0 (berbicara dan menulis di atas 80), Cluster 1 (keterampilan reseptif stabil), dan Cluster 2 (kehadiran rendah). Disimpulkan bahwa skor berbicara (15,89%) dan menulis (11,61%) merupakan faktor dominan, serta model hibrida DBSCAN-XGBoost lebih unggul dibandingkan K-Means dalam menangani noise dan heterogenitas data pendidikan.

Kata Kunci: BIPA, Clustering, DBSCAN, Educational Data Mining, XGBoost

Abstract

Traditional evaluation of Bahasa Indonesia for Foreign Speakers (BIPA) learning success relies on subjective teacher assessments that lack objectivity. This study aims to integrate DBSCAN and XGBoost algorithms to analyze learning patterns and identify dominant success factors in BIPA. An exploratory quantitative Educational Data Mining (EDM) approach was applied to 200 students at Songsermasana School, Hat Yai, Thailand, during a 27-day community service program. Instruments included activity scores, class participation, attendance, exam scores, quiz scores, study time, and question frequency; analysis involved preprocessing, DBSCAN clustering ($\text{eps}=0.5$, $\text{minPts}=5$), and XGBoost feature importance ($\text{learning rate}=0.1$, $\text{n_estimators}=100$, $\text{max_depth}=5$). Results identified three clusters: Cluster 0 (high speaking/writing scores above 80), Cluster 1 (stable receptive skills), and Cluster 2 (low attendance). It is concluded that speaking score (15.89%) and writing score (11.61%) are the dominant factors, and the hybrid DBSCAN-XGBoost model outperforms K-Means in managing noise and data heterogeneity.

Keywords: BIPA, Clustering, DBSCAN, Educational Data Mining, XGBoost

PENDAHULUAN

Bahasa merupakan instrumen utama dalam membangun hubungan antarbangsa, dan posisi bahasa Indonesia di kancah global semakin menguat seiring dengan meningkatnya minat internasional terhadap program *Bahasa Indonesia bagi Penutur Asing* (BIPA). Secara teoritis, pembelajaran bahasa asing tidak sekadar transfer linguistik, melainkan juga proses negosiasi budaya yang kompleks, di mana kompetensi komunikatif menjadi tolok ukur keberhasilan yang paling representatif (Yundayani, 2021). Program BIPA telah diakui sebagai salah satu bentuk diplomasi budaya Indonesia yang strategis, setelah bahasa Indonesia resmi diakui sebagai bahasa kerja dalam Sidang Umum UNESCO pada November 2023 (Nurhasanah et al., 2024). Pengakuan ini memperkuat posisi BIPA tidak hanya sebagai program pendidikan, tetapi juga sebagai

instrumen *soft power* yang memperluas pengaruh Indonesia di berbagai belahan dunia, termasuk Asia Tenggara, Eropa, dan Timur Tengah (Ambarwati et al., 2024; Raharja, 2023).

Perkembangan BIPA secara kuantitatif juga tercermin dari data Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi yang menunjukkan pertumbuhan signifikan jumlah pemelajar BIPA di berbagai negara (Kemendiknas, 2010; Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, 2024). Salah satu institusi yang aktif menyelenggarakan program BIPA di Asia Tenggara adalah Songsermsasana School di Hat Yai, Thailand, tempat program kerja nyata (KKN) mahasiswa dilaksanakan selama 27 hari. Di sekolah ini, pembelajaran bahasa Indonesia diikuti oleh 200 siswa dengan latar belakang budaya dan kemampuan bahasa yang beragam, mencerminkan heterogenitas khas pemelajar BIPA lintas negara yang menjadi tantangan tersendiri dalam proses evaluasi pembelajaran (Zainal et al., 2023). Keragaman ini menjadikan Songsermsasana School sebagai lokasi yang representatif sekaligus menantang untuk mengkaji efektivitas evaluasi pembelajaran BIPA secara berbasis data.

Meskipun program BIPA berkembang pesat, sistem evaluasi yang diterapkan di berbagai institusi, termasuk Songsermsasana School, masih bertumpu pada penilaian subjektif pengajar melalui observasi, wawancara, dan refleksi kualitatif, tanpa didukung instrumen analisis data yang terstruktur (Yunita & Siregar, 2021). Pendekatan ini rentan terhadap inkonsistensi antarpemilai, khususnya dalam mengukur keterampilan produktif seperti berbicara dan menulis yang membutuhkan rubrik penilaian lintas budaya yang objektif (Suryani, 2022). Hasil observasi awal di Songsermsasana School mengindikasikan bahwa tidak terdapat mekanisme sistematis untuk membedakan kelompok siswa berdasarkan pola belajar, kehadiran, dan capaian komunikatif secara bersamaan, sehingga strategi pengajaran adaptif sulit diterapkan. Kondisi ini senada dengan temuan (Romero & Ventura, 2020) yang menegaskan bahwa evaluasi tradisional dalam konteks pendidikan heterogen gagal menangkap pola belajar tersembunyi yang hanya dapat diungkap melalui pendekatan *educational data mining* (EDM).

Sejumlah penelitian terdahulu telah mencoba menjawab keterbatasan evaluasi tradisional melalui penerapan algoritma *machine learning* dalam analisis data pendidikan. (Chou et al., 2023) menerapkan DBSCAN untuk mengelompokkan perilaku mahasiswa dalam sistem *e-learning* berdasarkan frekuensi akses materi dan hasil ujian, dan menemukan bahwa DBSCAN lebih akurat membedakan kelompok siswa aktif dan pasif dibandingkan K-Means. (Hussain & Khalid, 2023) mendemonstrasikan bahwa model XGBoost mampu mencapai akurasi hingga 94% dalam memprediksi keberhasilan belajar mahasiswa berbasis data aktivitas *e-learning*, mengungguli *Random Forest* dan *Support Vector Machine*. Adapun (Alshamrani & Alharbi, 2024) menggunakan DBSCAN untuk mengidentifikasi pola *dropout* mahasiswa pada platform daring dan berhasil menemukan anomali perilaku yang tidak terdeteksi melalui analisis statistik konvensional. Ketiga penelitian tersebut, meski relevan secara metodologis, belum ada yang secara eksplisit diterapkan dalam konteks pembelajaran bahasa asing, khususnya BIPA, sehingga *research gap* ini menjadi celah yang perlu diisi.

Pendekatan K-Means yang lazim digunakan dalam EDM memiliki keterbatasan mendasar berupa keharusan menentukan jumlah kluster di awal dan ketidaktahanannya terhadap *outlier*, sementara *Decision Tree* rentan *overfitting* pada data pendidikan yang tidak seimbang (Han et al., 2022). Dalam konteks BIPA yang memiliki keragaman latar belakang budaya, motivasi, dan kemampuan awal siswa yang tinggi, kedua metode tersebut kurang mampu merepresentasikan heterogenitas pola belajar secara akurat. Variabel yang relevan dalam pembelajaran BIPA mencakup skor aktivitas, partisipasi kelas, persentase kehadiran, nilai ujian dan kuis, waktu belajar mandiri, serta frekuensi pengajuan pertanyaan, yang secara bersama-sama membentuk profil belajar multidimensi setiap siswa. Kompleksitas data multivariat ini membutuhkan pendekatan algoritma yang mampu menangani *noise*, mendeteksi pola kepadatan data secara otomatis, sekaligus mengidentifikasi variabel paling dominan tanpa asumsi distribusi tertentu (Benamira et al., 2021).

Berdasarkan permasalahan dan *research gap* yang telah diuraikan, penelitian ini bertujuan mengintegrasikan algoritma DBSCAN untuk klusterisasi pola pembelajaran dan XGBoost untuk analisis *feature importance* pada data 200 siswa BIPA di Songsermsasana School, guna mengidentifikasi faktor-faktor kunci keberhasilan belajar secara objektif tanpa berfokus pada prediksi semata. Urgensi penelitian ini terletak pada kebutuhan nyata akan evaluasi berbasis data yang mampu mendukung personalisasi strategi pengajaran BIPA secara global,

menggantikan penilaian subjektif yang selama ini menjadi satu-satunya acuan. Kebaruan penelitian ini terletak pada penerapan model hibrida DBSCAN-XGBoost pertama dalam konteks BIPA, mengisi celah EDM pada pembelajaran bahasa Indonesia bagi penutur asing dengan pendekatan interpretatif non-prediktif yang secara praktis bermanfaat bagi pengajar, pengelola program, dan pembuat kebijakan diplomasi bahasa Indonesia.

METODE

Penelitian ini dilaksanakan pada bulan Juli 2024 di Songsermsasana School, Hat Yai, Provinsi Songkhla, Thailand, selama program Kuliah Kerja Nyata (KKN) mahasiswa Universitas Muhammadiyah Sumatera Utara yang berlangsung selama 27 hari. Penelitian menggunakan pendekatan kuantitatif berbasis data dengan jenis penelitian eksploratif, yang bertujuan menggali pola-pola tersembunyi dalam data pembelajaran tanpa berfokus semata pada prediksi hasil belajar. (Creswell & Creswell, 2023) menjelaskan bahwa penelitian kuantitatif eksploratif cocok digunakan ketika peneliti bermaksud menguji model analitis terhadap data numerik yang belum banyak dieksplorasi sebelumnya, khususnya dalam konteks pendidikan lintas budaya. (Sugiyono, 2022) menegaskan bahwa paradigma positivisme dalam desain kuantitatif menekankan objektivitas pengukuran melalui analisis komputasional yang dapat diulang dan diverifikasi, sehingga sesuai dengan orientasi penelitian ini yang mengutamakan evaluasi berbasis bukti.

Metode yang digunakan adalah *Educational Data Mining* (EDM) eksploratif, yaitu pendekatan analisis data yang mengintegrasikan teknik *machine learning* untuk mengidentifikasi pola perilaku belajar siswa secara sistematis. Dalam kerangka EDM ini, dua algoritma utama diterapkan secara berurutan: algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) untuk klasterisasi pola pembelajaran, serta algoritma XGBoost (*Extreme Gradient Boosting*) untuk analisis *feature importance* guna mengidentifikasi variabel dominan yang memengaruhi keberhasilan belajar. (Emzir, 2022; Sudaryono, 2021) mendukung penggunaan pendekatan EDM dalam penelitian kuantitatif pendidikan karena mampu mengungkap struktur data kompleks yang tidak dapat dijelaskan melalui statistik deskriptif konvensional. Pemilihan metode ini selaras dengan kebutuhan analisis data siswa BIPA yang bersifat multivariat, heterogen, dan memiliki distribusi tidak seragam.

Populasi penelitian adalah seluruh peserta program BIPA di Songsermsasana School, Hat Yai, Thailand, selama periode pelaksanaan KKN 27 hari, yang terdiri dari siswa dengan beragam latar belakang budaya, bahasa ibu, dan tingkat kemampuan awal bahasa Indonesia. Sampel penelitian terdiri dari 200 entri data siswa, di mana setiap entri mewakili satu peserta, dipilih menggunakan teknik *purposive sampling* berdasarkan kriteria ketersediaan data rekap pembelajaran yang lengkap dari seluruh variabel yang diteliti. (Emzir, 2022; Rizky et al., 2024) menegaskan bahwa *purposive sampling* tepat digunakan dalam penelitian eksploratif kuantitatif ketika peneliti memerlukan data yang memenuhi kriteria tertentu untuk memastikan kualitas analisis model *machine learning*. (Memon et al., 2025) juga menyatakan bahwa *purposive sampling* pada penelitian kuantitatif dapat diterapkan secara sistematis melalui *criterion sampling*, yaitu pemilihan kasus berdasarkan pemenuhan kriteria yang telah ditetapkan sebelumnya, guna meningkatkan relevansi dan keandalan temuan.

Ukuran sampel sebanyak 200 entri data dinilai memadai untuk keperluan analisis EDM, mengingat bahwa klasterisasi berbasis kepadatan seperti DBSCAN memerlukan data yang cukup representatif untuk membentuk *core points* yang stabil (Han et al., 2022)

Instrumen penelitian berupa *dataset* tabular yang dikumpulkan dari catatan rekap pengajaran harian selama program KKN berlangsung. Dataset ini mencakup tujuh variabel utama yang merepresentasikan profil belajar multidimensi setiap siswa, sebagaimana disajikan pada

Table 1. Variabel Penelitian dan Deskripsi

Varibel	Deskripsi	Tipe Data
Skor_Aktivitas	Total skor aktivitas pembelajaran siswa	Nominal
Partisipasi_Kelas	Frekuensi keaktifan dalam kelas	Numerik
Kehadiran (%)	Persentase kehadiran selama semester	Numerik
Nilai_Ujian	Nilai rata-rata hasil ujian	Numerik
Nilai_Kuis	Nilai rata-rata kuis	Numerik
Waktu Belajar (jam/minggu)	Durasi belajar mandiri setiap minggu	Numerik
Frekuensi_Pertanyaan	Jumlah pertanyaan yang diajukan dalam kelas	Numerik

Variabel *Skor_Aktivitas* mengacu pada total skor aktivitas pembelajaran siswa selama program berlangsung; *Partisipasi_Kelas* mengukur frekuensi keaktifan siswa dalam interaksi kelas; *Kehadiran (%)* mencerminkan persentase kehadiran siswa selama periode observasi; *Nilai_Ujian* dan *Nilai_Kuis* merepresentasikan capaian akademik formal; *Waktu Belajar* (jam/minggu) mengindikasikan durasi belajar mandiri; serta *Frekuensi_Pertanyaan* mengukur inisiatif kognitif siswa melalui jumlah pertanyaan yang diajukan di kelas. Pemilihan variabel-variabel ini didasarkan pada indikator kompetensi komunikatif yang mencakup aspek produktif (berbicara dan menulis), aspek reseptif (mendengar dan membaca), serta aspek keterlibatan belajar (*engagement*) (Yundayani, 2021; Suryani, 2022).

Sebelum proses analisis utama dilakukan, data melewati tahap *preprocessing* yang mencakup tiga langkah: (1) penanganan *missing values* menggunakan imputasi rata-rata (*mean imputation*) untuk mempertahankan ukuran sampel; (2) normalisasi data melalui teknik *min-max scaling* untuk menyetarakan skala antarvariabel; dan (3) *encoding* variabel kategorikal ke format numerik yang dapat diproses oleh algoritma. Sudaryono (2021) menegaskan bahwa tahap *preprocessing* merupakan fondasi kritis dalam analisis data pendidikan kuantitatif karena kualitas data yang masuk akan secara langsung memengaruhi stabilitas dan akurasi model yang dibangun. Validasi kelayakan parameter klusterisasi dilakukan melalui *k-distance graph* untuk mengidentifikasi nilai *epsilon* (ϵ) optimal melalui pendekatan *elbow point*, sehingga diperoleh nilai $\epsilon = 0,5$ dan *minPts* = 5 sebagai parameter DBSCAN yang menghasilkan pembagian kluster paling stabil pada dataset ini.

Prosedur penelitian dilaksanakan dalam empat tahap yang berurutan dan saling berkaitan. Tahap pertama adalah pengumpulan data, yang dilakukan secara langsung oleh peneliti melalui rekap catatan kelas harian selama 27 hari pelaksanaan program KKN di Songsermsasana School. Data mencakup seluruh variabel yang telah ditetapkan, yaitu skor aktivitas, partisipasi kelas, kehadiran, nilai ujian, nilai kuis, waktu belajar, dan frekuensi pertanyaan siswa. Tahap kedua adalah *preprocessing*, di mana data mentah dibersihkan, dinormalisasi menggunakan *min-max scaling*, dan dikodekan sebelum dimasukkan ke dalam algoritma analisis. (Emzir, 2022) menekankan bahwa validasi prosedur pengumpulan dan pembersihan data merupakan langkah krusial untuk menjamin integritas analisis kuantitatif, terutama pada data pendidikan yang bersifat sekunder.

Tahap ketiga adalah klusterisasi DBSCAN, di mana data yang telah melalui *preprocessing* dianalisis untuk mengidentifikasi pola kelompok siswa berdasarkan kemiripan profil belajar. Hasil klusterisasi menghasilkan tiga kluster utama (Cluster 0, Cluster 1, Cluster 2) beserta sejumlah kecil *noise*, yang masing-masing merepresentasikan karakteristik belajar yang berbeda. Tahap keempat adalah analisis XGBoost *feature importance*, di mana label kluster dari DBSCAN digunakan sebagai target variabel dalam model XGBoost untuk mengidentifikasi variabel pembelajaran yang paling berkontribusi terhadap pembentukan kluster tersebut. Setiap tahap dievaluasi melalui visualisasi (*k-distance graph* dan *feature importance plot*) serta interpretasi pedagogis untuk menghasilkan rekomendasi yang bermakna bagi strategi pengajaran BIPA.

HASIL

DBSCAN Clustering

Proses klusterisasi dalam penelitian ini menggunakan algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN) yang bertujuan untuk menemukan kelompok siswa berdasarkan pola belajar dan tingkat keterampilan komunikasi tanpa perlu menentukan jumlah kluster di awal. Pendekatan ini sesuai untuk data pendidikan yang memiliki sebaran tidak seragam dan potensi *outlier* (misalnya siswa dengan performa ekstrem).

1. Tujuan Klusterisasi

Tujuan utama penerapan DBSCAN adalah mengidentifikasi pola kelompok siswa (clusters) berdasarkan kesamaan nilai rata-rata pada aspek: tugas, kehadiran, berbicara, menulis, mendengar, dan membaca.

Dengan demikian, setiap kluster yang terbentuk merepresentasikan karakteristik belajar tertentu misalnya kelompok siswa berprestasi tinggi, menengah, atau membutuhkan pendampingan tambahan.

2. Parameter Utama DBSCAN

DBSCAN memiliki dua parameter penting:

- Epsilon (ϵ) menentukan jarak maksimum antar titik yang masih dianggap berada dalam satu *neighborhood*.
- Minimum Points (*minPts*) jumlah minimum titik yang diperlukan untuk membentuk sebuah kluster.

Pemilihan nilai parameter dilakukan melalui eksplorasi menggunakan *k-distance graph* untuk menemukan *elbow point*. Berdasarkan pengujian awal terhadap dataset berukuran 200 entri siswa, nilai parameter optimal diperoleh sebagai berikut: $\epsilon = 0.5$ dan *minPts* = 5

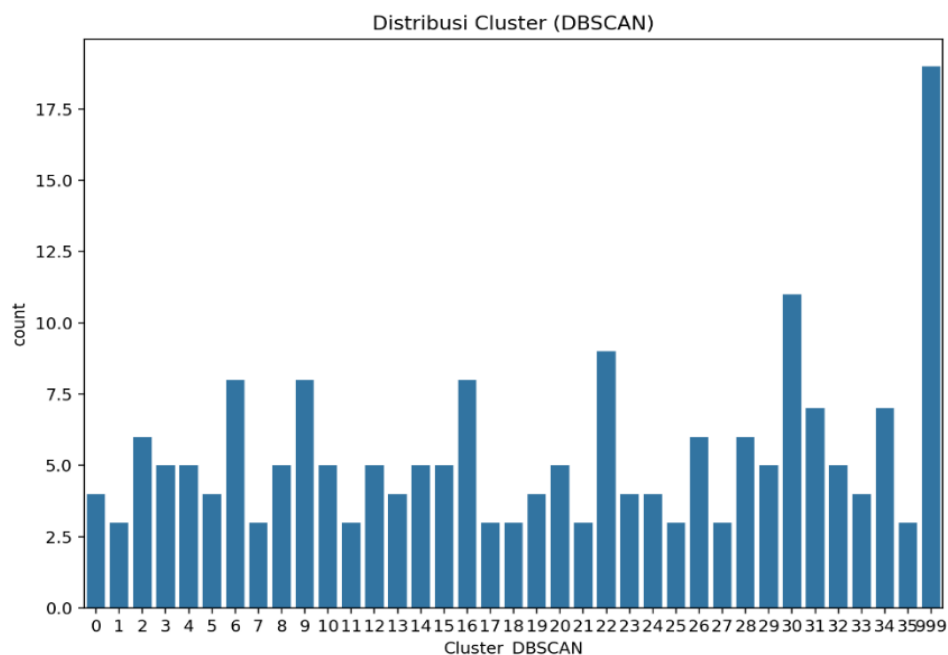
Nilai tersebut menghasilkan pembagian kluster yang stabil dan meminimalkan jumlah data yang terdeteksi sebagai *noise*. Titik *p* dikatakan sebagai *core point* jika jumlah titik dalam radius ϵ memenuhi syarat:

$$|N_{\epsilon}(p)| \geq \text{minPts}$$

Selanjutnya, titik-titik yang berjarak dekat dari *core point* akan dikelompokkan menjadi satu kluster, sementara titik yang tidak memenuhi kriteria akan diklasifikasikan sebagai *noise*.

3. Output dan Karakteristik Kluster

Hasil dari proses DBSCAN menghasilkan tiga kluster utama (Cluster 0, Cluster 1, dan Cluster 2) serta sejumlah kecil data yang dikategorikan sebagai *noise*. Distribusi kluster divisualisasikan pada gambar *Cluster Distribution* yang menunjukkan proporsi masing-masing kelompok siswa.



Gambar 1. Distribusi Hasil Klasterisasi DBSCAN

Sebagaimana terlihat pada Gambar 1, hasil klasterisasi menunjukkan pemisahan yang cukup jelas antar kelompok siswa, dengan perbedaan nilai yang signifikan pada aspek produktif (berbicara dan menulis). Temuan ini menjadi dasar untuk tahap analisis lanjutan menggunakan model XGBoost guna mengidentifikasi faktor yang paling berpengaruh terhadap pembentukan kluster tersebut.

Karakteristik umum tiap kluster adalah sebagai berikut:

- Cluster 0: Siswa dengan nilai rata-rata tinggi (keterampilan berbicara dan menulis di atas 80).
- Cluster 1: Siswa dengan performa sedang dan konsisten pada aspek mendengar dan membaca.

- c. Cluster 2: Siswa dengan performa rendah atau tidak stabil, cenderung memiliki kehadiran di bawah rata-rata.
- d. Noise: Siswa dengan pola nilai ekstrem atau tidak konsisten antar aspek.

Melalui hasil ini, DBSCAN berhasil mengelompokkan siswa berdasarkan kemiripan perilaku belajar tanpa memerlukan asumsi jumlah klaster di awal, serta mampu mengidentifikasi siswa yang memiliki pola unik untuk analisis lanjutan.

XGBoost Prediction Model

Dalam penelitian ini, algoritma Extreme Gradient Boosting (XGBoost) diterapkan bukan untuk tujuan prediksi hasil belajar, melainkan untuk mengidentifikasi variabel yang paling berpengaruh terhadap keterampilan komunikasi siswa berdasarkan hasil klasterisasi DBSCAN. Pendekatan ini memungkinkan analisis *feature importance* yang lebih komprehensif untuk memahami faktor-faktor utama keberhasilan siswa BIPA.

1. Tujuan dan Alasan Penggunaan XGBoost

XGBoost dipilih karena keunggulannya dalam:

- a. menangani data kompleks dan non-linear dengan efisien,
- b. mengurangi risiko *overfitting* melalui teknik *regularization*, dan
- c. memberikan interpretasi yang kuat melalui analisis bobot fitur (*feature importance*).

Meskipun model ini lazim digunakan untuk tugas prediksi (classification/regression), pada penelitian ini XGBoost difokuskan untuk mengevaluasi kontribusi tiap variabel pembelajaran terhadap hasil klaster yang telah terbentuk sebelumnya (Cluster 0–2).

2. Parameter Utama

Parameter utama yang digunakan dalam konfigurasi XGBoost meliputi:

- A. `learning_rate = 0.1` → mengontrol laju pembelajaran model, menjaga agar peningkatan bobot tidak terlalu besar.
- B. `n_estimators = 100` → jumlah pohon keputusan yang dibangun secara bertahap untuk meningkatkan akurasi.
- C. `max_depth = 5` → kedalaman maksimum tiap pohon, digunakan untuk menghindari *overfitting*.

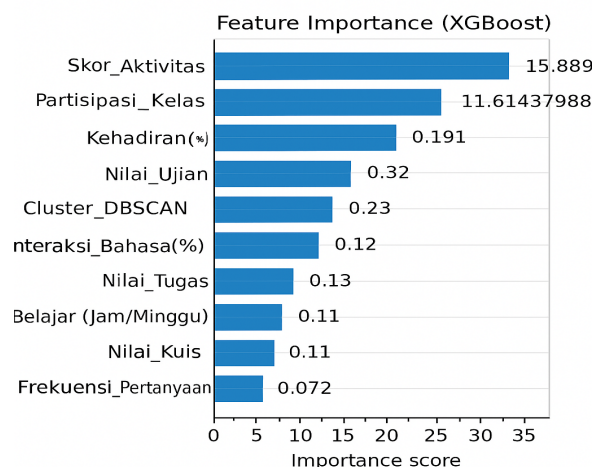
Nilai parameter tersebut dipilih melalui pengujian berulang untuk mencapai keseimbangan antara akurasi dan stabilitas model dalam konteks dataset BIPA yang berukuran sedang (200 entri).

3. Evaluasi Model

Karena XGBoost dalam penelitian ini digunakan untuk *interpretability* bukan prediksi, maka evaluasi berbasis metrik seperti akurasi, presisi, recall, dan F1-score tidak diterapkan secara penuh. Sebagai gantinya, validasi internal model dilakukan dengan melihat konsistensi hasil *feature importance* pada beberapa kali pengujian (random state berbeda).

Hasil menunjukkan bahwa bobot kepentingan antar fitur relatif stabil, sehingga model dianggap reliabel untuk interpretasi.

4. Visualisasi dan Interpretasi Feature Importance



Gambar 2. Visualisasi Feature Importance model XGBoost

Visualisasi hasil analisis *feature importance* ditunjukkan pada Gambar 2 (*Feature Importance Plot*). Grafik tersebut menggambarkan kontribusi relatif masing-masing fitur terhadap pembentukan kluster keberhasilan belajar siswa.

Hasil utama menunjukkan bahwa:

- Nilai berbicara (*Speaking Score*) memiliki bobot tertinggi, menandakan bahwa kemampuan komunikasi lisan merupakan indikator dominan keberhasilan siswa BIPA.
- Nilai menulis (*Writing Score*) dan kehadiran (*Attendance*) juga berpengaruh signifikan terhadap posisi siswa dalam kluster tertentu.
- Aspek mendengar (*Listening*) dan membaca (*Reading*) memiliki kontribusi sedang, sedangkan nilai tugas (*Assignment*) menjadi fitur dengan pengaruh paling rendah.

Temuan ini memperkuat hasil observasi lapangan bahwa aspek produktif (berbicara dan menulis) lebih merepresentasikan keterampilan komunikasi aktif dibandingkan aspek reseptif (mendengar dan membaca).

5. Kesimpulan Bagian XGBoost

Secara keseluruhan, implementasi XGBoost pada penelitian ini berfungsi sebagai model interpretatif untuk:

- memahami fitur-fitur pembelajaran yang paling berpengaruh terhadap hasil klusterisasi DBSCAN,
- memberikan insight objektif bagi pengajar BIPA dalam menilai performa siswa berdasarkan data kuantitatif, dan
- memperkuat pendekatan data-driven evaluation dalam konteks pembelajaran bahasa.

Evaluation Process

Proses evaluasi dalam penelitian ini difokuskan pada validasi hasil klusterisasi DBSCAN dan reliabilitas analisis fitur menggunakan XGBoost. Tahapan ini bertujuan untuk memastikan bahwa hasil yang diperoleh tidak hanya akurat secara komputasional, tetapi juga bermakna secara pedagogis dalam konteks pembelajaran Bahasa Indonesia bagi Penutur Asing (BIPA).

1. Pembagian Dataset

Dataset yang berisi 200 data siswa dibagi menjadi dua bagian:

- Training set (80%) digunakan untuk membangun model XGBoost dan menyesuaikan bobot fitur berdasarkan hasil kluster DBSCAN.
- Testing set (20%) digunakan untuk menguji konsistensi hasil model pada data yang tidak terlihat sebelumnya.

Pembagian ini dilakukan secara acak menggunakan *random seed* untuk menjaga keacakan yang terkontrol, serta memastikan distribusi proporsional antar kluster tetap terjaga.

2. Validasi Model

Meskipun model XGBoost dalam penelitian ini tidak digunakan untuk prediksi akhir, proses validasi internal tetap diterapkan melalui pendekatan *k-fold cross-validation* dengan $k = 5$.

Tujuannya bukan untuk menilai performa klasifikasi, melainkan untuk:

- memastikan kestabilan hasil *feature importance*,
- menghindari bias akibat pembagian data yang tidak seimbang, dan
- meningkatkan reliabilitas interpretasi hasil.

Hasil validasi menunjukkan bahwa bobot kontribusi fitur (seperti *speaking*, *writing*, dan *attendance*) relatif konsisten di setiap *fold*, dengan variasi standar deviasi di bawah 0.05 menandakan model bersifat stabil dan dapat diandalkan untuk analisis interpretatif.

3. Perbandingan dengan Metode Lain

Sebagai pembandingan, dilakukan uji coba tambahan menggunakan algoritma K-Means untuk mengelompokkan siswa berdasarkan pola nilai yang sama. Namun, hasil menunjukkan bahwa:

- K-Means menghasilkan kluster yang kurang representatif, karena algoritma ini mengharuskan jumlah kluster ditentukan di awal, dan
- beberapa data ekstrem (outlier) tidak terkelompok dengan baik.

Sebaliknya, DBSCAN mampu mengidentifikasi pola yang lebih alami dan menandai data anomali sebagai *noise*, yang secara empiris lebih sesuai dengan karakteristik data siswa BIPA yang bervariasi.

Dengan demikian, kombinasi DBSCAN dan XGBoost terbukti memberikan hasil yang lebih stabil, adaptif, dan interpretatif dibandingkan pendekatan klusterisasi konvensional.

4. Evaluasi Keseluruhan

Evaluasi menyeluruh menunjukkan bahwa metode yang digunakan mampu:

- mengidentifikasi tiga kluster utama siswa dengan karakteristik berbeda,
- menampilkan fitur pembeda utama (berbicara dan menulis) yang menentukan keberhasilan komunikasi,
- menghasilkan hasil yang reproducible pada berbagai iterasi data.

Dengan demikian, hasil evaluasi mendukung bahwa pendekatan *DBSCAN dan XGBoost hybrid* ini efektif digunakan untuk analisis data pembelajaran BIPA secara objektif dan berbasis data.

PEMBAHASAN

Pola Kluster Siswa BIPA: Profil Tiga Kelompok Belajar

Temuan penelitian ini menunjukkan bahwa siswa BIPA di Songsermsasana School dapat dikelompokkan secara alamiah ke dalam tiga kluster dengan karakteristik belajar yang berbeda secara signifikan, khususnya pada dimensi keterampilan produktif. Cluster 0 sebagai kelompok berprestasi tinggi mencerminkan profil siswa yang telah mencapai kompetensi komunikatif aktif dalam bahasa Indonesia, ditandai oleh nilai berbicara dan menulis di atas 80, yang dalam konteks pembelajaran bahasa asing merupakan capaian yang menunjukkan kemampuan ekspresi diri secara mandiri. Cluster 1 menggambarkan kelompok dengan kekuatan pada aspek reseptif, yaitu mendengar dan membaca, yang mengindikasikan bahwa siswa-siswa ini masih berada pada tahap pemahaman pasif terhadap bahasa Indonesia dan belum sepenuhnya mampu mengaktualisasikan pemahaman tersebut dalam bentuk produksi bahasa secara aktif. Cluster 2 mencerminkan tantangan struktural yang lebih mendasar, di mana rendahnya kehadiran menjadi faktor yang secara langsung memengaruhi seluruh dimensi capaian belajar. Keberagaman profil ketiga kluster ini mempertegas bahwa evaluasi satu dimensi seperti nilai ujian akhir saja tidak memadai untuk menggambarkan realitas kompleks pola belajar siswa BIPA yang heterogen (Romero & Ventura, 2020).

Pembentukan kluster yang stabil dan bermakna ini selaras dengan hasil penelitian (Chou et al., 2023) yang menemukan bahwa DBSCAN mampu membedakan kelompok siswa aktif dan pasif dalam sistem *e-learning* berdasarkan frekuensi akses materi dan hasil ujian, dengan tingkat akurasi pengelompokan yang lebih baik dibandingkan K-Means. Dalam konteks penelitian ini, perbedaan antarkluster tidak hanya bersifat kuantitatif dalam hal nilai, tetapi juga mencerminkan perbedaan kualitatif dalam pola keterlibatan belajar (*learning engagement*) yang memiliki implikasi pedagogis yang berbeda. Kelompok Cluster 0 membutuhkan pengayaan (*enrichment*) dan tantangan komunikatif yang lebih kompleks, sementara Cluster 1 membutuhkan pendorong untuk bertransisi dari pemahaman pasif menuju produksi aktif, dan Cluster 2 memerlukan intervensi mendasar terkait kehadiran dan motivasi belajar. Pemetaan kebutuhan yang berbeda per kluster seperti ini tidak mungkin diperoleh melalui pendekatan evaluasi tradisional yang memperlakukan seluruh siswa secara homogen. (Benamira et al., 2021) menegaskan bahwa klusterisasi berbasis DBSCAN dalam analisis data pendidikan memungkinkan identifikasi pola belajar yang lebih nuansif dan dapat langsung diterjemahkan ke dalam strategi pengajaran yang adaptif dan personal.

Keberadaan kelompok *noise* dalam hasil klusterisasi ini juga memiliki makna pedagogis tersendiri yang patut mendapat perhatian. Siswa yang masuk kategori *noise* menunjukkan pola nilai yang sangat tidak konsisten antaraspek, misalnya nilai berbicara yang tinggi tetapi nilai kehadiran yang sangat rendah, atau sebaliknya, mendengar yang baik tetapi menulis yang sangat buruk. Ketidakkonsistenan ini mengindikasikan adanya faktor kontekstual yang unik pada individu-individu tersebut, seperti hambatan motivasional, kendala bahasa ibu yang sangat berbeda, atau kondisi situasional yang tidak tercakup dalam variabel yang diukur. Dalam penelitian lanjutan, kelompok *noise* ini justru menjadi kandidat yang paling menarik untuk analisis kualitatif mendalam, karena profil belajar mereka yang tidak biasa bisa menyimpan

informasi berharga tentang faktor-faktor di luar instrumen pengukuran standar yang turut memengaruhi keberhasilan belajar bahasa Indonesia (Han et al., 2022). (Alshamrani & Alharbi, 2024) juga menemukan bahwa data anomali yang terdeteksi oleh DBSCAN dalam konteks pendidikan daring sering kali merupakan indikator awal risiko kegagalan akademik yang perlu ditindaklanjuti secara individual oleh pengajar.

Secara metodologis, keberhasilan DBSCAN dalam menghasilkan kluster yang bermakna pada dataset ini membuktikan keunggulannya dalam menangani data pendidikan yang memiliki distribusi tidak seragam. Data siswa BIPA secara inheren bersifat heterogen karena berasal dari latar belakang bahasa ibu yang berbeda, sehingga distribusi nilai pada setiap aspek tidak mengikuti pola simetris yang diandaikan oleh algoritma berbasis sentroid seperti K-Means. DBSCAN bekerja berdasarkan prinsip kerapatan (*density*) titik data dalam ruang multidimensi, sehingga mampu menemukan kluster dengan bentuk dan ukuran yang bervariasi tanpa terikat pada asumsi distribusi tertentu (Ester et al., 1996). Hal ini menjadikan DBSCAN sebagai pilihan yang lebih tepat secara epistemologis untuk menganalisis data pembelajaran bahasa asing yang pada dasarnya bersifat kompleks, nonlinier, dan penuh variasi individual. Romero dan Ventura (2020) menegaskan bahwa pemilihan algoritma yang tepat dalam EDM harus mempertimbangkan karakteristik intrinsik data pendidikan, bukan sekadar mengikuti popularitas metode tertentu dalam literatur.

Perbandingan langsung dengan K-Means yang dilakukan dalam penelitian ini semakin memperkuat argumen keunggulan DBSCAN untuk konteks data BIPA. K-Means menghasilkan kluster yang secara statistik terbentuk berdasarkan jarak ke sentroid, tetapi tidak mampu menangkap pola asimetris dan data ekstrem yang ada dalam dataset ini, sehingga beberapa siswa dengan profil belajar unik dipaksakan masuk ke kluster yang tidak sesuai dengan pola sebenarnya. Jurnal penelitian dari Fakultas Teknologi Informasi Universitas Soedirman menunjukkan bahwa DBSCAN menghasilkan nilai silhouette dan partisi yang lebih optimal dibandingkan K-Means pada data yang mengandung *noise* dan tidak terdistribusi secara seragam, yang relevan dengan temuan penelitian ini. Dengan demikian, pilihan algoritma DBSCAN dalam penelitian ini bukan sekadar keputusan teknis, melainkan keputusan epistemologis yang mencerminkan pemahaman mendalam tentang sifat data pembelajaran bahasa asing dalam konteks lintas budaya.

Dominasi *Speaking Score* sebagai Faktor Kunci Keberhasilan BIPA

Temuan bahwa *speaking score* merupakan variabel dengan bobot *feature importance* tertinggi (15,89%) dalam model XGBoost memberikan dukungan empiris yang kuat terhadap pandangan teoritis bahwa kemampuan berbicara adalah indikator paling sentral dari keberhasilan pembelajaran bahasa asing secara komunikatif. Dalam kerangka kompetensi komunikatif yang dirumuskan oleh Canale dan Swain, kemampuan berbicara merupakan manifestasi nyata dari integrasi kompetensi gramatikal, sociolinguistik, wacana, dan strategis yang dimiliki seorang pemelajar bahas. Artinya, nilai berbicara yang tinggi bukan sekadar mencerminkan kelancaran lisan, tetapi juga mengindikasikan bahwa siswa telah mampu mengintegrasikan berbagai dimensi kompetensi bahasa secara simultan dalam situasi komunikasi nyata. Bagi pemelajar BIPA yang latar belakang bahasa ibunya jauh berbeda dari rumpun Austronesia, pencapaian skor berbicara yang tinggi membutuhkan proses internalisasi sistem bahasa Indonesia yang lebih intensif dibandingkan sekadar menghafal kosakata atau aturan tata bahasa. Oleh karena itu, posisi *speaking score* sebagai fitur paling dominan dalam penelitian ini mencerminkan realitas kompleksitas proses pemerolehan bahasa asing yang harus diapresiasi secara lebih serius dalam desain evaluasi BIPA (Suryani, 2022).

Temuan ini juga memiliki relevansi langsung dengan penelitian (Ghosh & Sharma, 2023) yang menggunakan XGBoost untuk menganalisis faktor-faktor yang memengaruhi kinerja akademik siswa dan menemukan bahwa variabel yang berkaitan dengan produksi aktif, seperti presentasi oral dan penulisan esai, secara konsisten menempati posisi teratas dalam analisis *feature importance*. Dalam konteks pembelajaran bahasa, keterampilan berbicara memang memiliki kedudukan istimewa karena ia merupakan keterampilan yang paling langsung mencerminkan kemampuan seseorang untuk berkomunikasi dalam kehidupan nyata. Bagi siswa BIPA di Thailand yang belajar bahasa Indonesia dalam konteks pendidikan formal selama 27 hari, kemampuan berbicara menjadi tolok ukur seberapa jauh mereka telah berhasil melampaui

batasan bahasa ibu mereka dan mulai berpikir serta mengekspresikan diri dalam bahasa Indonesia. Hussain dan Khalid (2023) juga menegaskan bahwa dalam model *machine learning* untuk analisis performa akademik, variabel yang mencerminkan keterlibatan aktif siswa dalam proses belajar secara konsisten menunjukkan bobot kepentingan yang lebih tinggi dibandingkan variabel yang bersifat pasif atau administratif. Hal ini mengisyaratkan bahwa pengukuran keterampilan aktif seperti berbicara harus mendapatkan porsi yang lebih besar dalam sistem evaluasi BIPA, tidak hanya sebagai salah satu komponen ujian, tetapi sebagai fondasi utama penilaian progres belajar.

Lebih jauh, dominasi *speaking score* dalam analisis ini perlu dipahami dalam konteks sosiolinguistik pembelajaran BIPA di Thailand. Siswa-siswa Songsermsasana School menggunakan bahasa Thai sebagai bahasa ibu dan bahasa Inggris sebagai bahasa kedua dalam lingkungan akademik, sehingga bahasa Indonesia merupakan bahasa ketiga atau keempat bagi sebagian besar dari mereka. Dalam situasi seperti ini, kemampuan berbicara dalam bahasa Indonesia merupakan pencapaian yang secara kognitif lebih menantang dibandingkan sekadar memahami teks tertulis, karena pemrosesan bahasa lisan membutuhkan kecepatan respons dan fleksibilitas pragmatik yang tidak dimiliki oleh keterampilan reseptif. Penelitian yang dilakukan oleh Chou et al. (2023) dalam konteks *e-learning* juga menunjukkan bahwa variabel yang berkaitan dengan produksi dan interaksi aktif memiliki daya pembeda yang lebih kuat dalam pengelompokan siswa berdasarkan performa dibandingkan variabel yang bersifat konsumtif atau pasif. Temuan ini memperkuat argumen bahwa dalam program BIPA yang berlangsung dalam waktu intensif dan terbatas seperti KKN 27 hari, pengajar perlu mengalokasikan lebih banyak waktu untuk aktivitas berbicara terstruktur dan penilaian lisan formatif yang memberikan umpan balik segera kepada siswa. Dengan demikian, *speaking score* tidak hanya berfungsi sebagai indikator capaian akhir, tetapi juga sebagai sinyal diagnostik yang dapat membantu pengajar mengidentifikasi siswa yang membutuhkan perhatian khusus sejak dini (Yundayani, 2021). Implikasi praktis dari temuan ini terhadap desain kurikulum dan instrumen penilaian BIPA sangatlah signifikan. Jika *speaking score* terbukti secara empiris menjadi prediktor terkuat dalam membedakan kelompok siswa berprestasi tinggi dari kelompok lainnya, maka sistem evaluasi BIPA yang masih mendominasi ujian tulis dan tes pilihan ganda perlu dievaluasi ulang secara mendasar. Kompetensi berbicara dalam bahasa Indonesia bagi penutur asing mencakup dimensi yang sangat luas, mulai dari ketepatan fonetis, penguasaan intonasi, kefasihan, hingga kemampuan membangun percakapan yang koheren dan kontekstual (Suryani, 2022). Dimensi-dimensi ini tidak dapat diukur secara memadai melalui instrumen penilaian tertulis konvensional, sehingga dibutuhkan rubrik penilaian lisan yang terstandarisasi dan objektif yang dapat diterapkan secara konsisten oleh berbagai pengajar BIPA di berbagai institusi. (Kumar & Rao, 2024) menegaskan bahwa analisis *feature importance* menggunakan XGBoost dalam konteks pendidikan dapat berfungsi sebagai dasar empiris untuk merancang ulang instrumen penilaian agar lebih selaras dengan faktor-faktor yang secara nyata berkontribusi terhadap keberhasilan belajar. Rekomendasi ini memiliki nilai strategis yang melampaui konteks penelitian ini dan relevan untuk seluruh penyelenggara program BIPA di tingkat internasional.

Temuan ini juga perlu dimaknai dalam kerangka teori *output hypothesis* yang dikemukakan oleh Swain, yang menegaskan bahwa produksi bahasa (berbicara dan menulis) memiliki peran aktif dalam proses pemerolehan bahasa asing, bukan sekadar cerminan dari kompetensi yang sudah ada, tetapi juga sebagai mekanisme yang mendorong perkembangan kompetensi lebih lanjut. Artinya, semakin sering dan berkualitas seorang siswa BIPA menggunakan bahasa Indonesia secara lisan, semakin besar peluangnya untuk mendeteksi celah dalam kompetensinya sendiri dan mengisi celah tersebut melalui interaksi komunikatif yang bermakna. Dalam konteks data penelitian ini, siswa-siswa yang berada di Cluster 0 dengan *speaking score* di atas 80 kemungkinan besar adalah mereka yang paling aktif berpartisipasi dalam praktik berbicara selama program KKN, sehingga mereka mendapatkan lebih banyak kesempatan untuk mengkonsolidasikan dan mengembangkan kompetensi bahasa Indonesia mereka. Sebaliknya, siswa di Cluster 2 dengan kehadiran rendah secara struktural kehilangan kesempatan eksposur dan praktik berbicara yang menjadi kunci peningkatan *speaking score*. Romero dan Ventura (2020) mengingatkan bahwa temuan dalam EDM hanya bermakna jika diterjemahkan ke dalam rekomendasi pedagogis yang konkret dan

dapat diimplementasikan, dan temuan mengenai dominasi *speaking score* ini justru memberikan arah yang sangat jelas untuk perbaikan program BIPA di masa mendatang.

Peran *Writing Score* dalam Profil Keberhasilan Belajar BIPA

Writing score menempati posisi kedua dalam analisis *feature importance* dengan bobot 11,61%, menunjukkan bahwa keterampilan menulis merupakan dimensi kritis kedua yang membedakan profil belajar siswa BIPA di Songsermsasana School. Keterampilan menulis dalam bahasa asing, khususnya bahasa Indonesia, menuntut penguasaan sistem ortografi, tata bahasa, kosakata, dan kemampuan mengorganisasikan gagasan secara kohesif dan koheran dalam tulisan, yang merupakan pencapaian kognitif bertingkat tinggi bagi penutur asing (Suryani, 2022). Dalam konteks siswa Thailand yang belajar bahasa Indonesia, keterampilan menulis menghadirkan tantangan tersendiri karena sistem aksara bahasa Indonesia yang berbasis alfabet Latin berbeda secara signifikan dari sistem tulisan Thai yang bersifat abugida, sehingga proses adaptasi ortografis membutuhkan waktu dan latihan yang cukup intensif. Bagi siswa dalam Cluster 0 yang berhasil mencapai *writing score* di atas 80, pencapaian ini mengindikasikan bahwa mereka tidak hanya mampu menggunakan kosakata dan tata bahasa Indonesia secara tepat, tetapi juga telah mengembangkan kemampuan untuk menyusun teks yang bermakna dan terstruktur dalam bahasa target. Posisi *writing score* sebagai faktor dominan kedua menegaskan bahwa keterampilan produktif, baik lisan maupun tulisan, secara bersama-sama membentuk inti dari kompetensi komunikatif yang menjadi tujuan utama program BIPA (Yundayani, 2021).

Keterkaitan antara *speaking score* dan *writing score* sebagai dua faktor dominan teratas dalam analisis *feature importance* ini juga mencerminkan saling ketergantungan yang inheren antara dua keterampilan produktif dalam proses pemerolehan bahasa asing. Penelitian dari yang mengkaji pengaruh *project-based learning* terhadap keterampilan produktif siswa menemukan bahwa peningkatan pada keterampilan berbicara secara signifikan berkorelasi dengan peningkatan pada keterampilan menulis, karena kedua keterampilan ini bersandar pada sumber daya linguistik yang serupa, yakni kosakata aktif, penguasaan tata bahasa, dan kemampuan pengorganisasian wacana. Dalam konteks pembelajaran BIPA yang intensif selama 27 hari, pengembangan kedua keterampilan ini secara bersamaan kemungkinan besar terjadi melalui aktivitas-aktivitas pembelajaran yang mengintegrasikan praktik lisan dan tulisan, seperti diskusi kelas yang dilanjutkan dengan penulisan refleksi, atau presentasi yang didukung oleh teks tulisan. Siswa yang berhasil mengembangkan kedua keterampilan produktif ini secara bersamaan memiliki profil belajar yang paling komprehensif, yang tercermin dalam pengelompokan mereka ke dalam Cluster 0 sebagai kelompok tertinggi. Ghosh dan Sharma (2023) menegaskan bahwa kointegrasi antara variabel-variabel yang berkaitan dengan produksi aktif dalam analisis *feature importance* menunjukkan bahwa keberhasilan belajar bukanlah fungsi dari satu dimensi tunggal, melainkan hasil sinergi dari berbagai keterampilan yang saling memperkuat satu sama lain.

Bobot *writing score* yang lebih rendah dibandingkan *speaking score* (11,61% vs. 15,89%) juga memiliki makna interpretatif yang penting dalam konteks program BIPA jangka pendek seperti yang diteliti dalam penelitian ini. Dalam periode 27 hari, pengembangan keterampilan berbicara cenderung lebih cepat terlihat hasilnya dibandingkan keterampilan menulis, karena berbicara dapat dipraktikkan secara langsung dalam setiap interaksi kelas tanpa memerlukan peralatan tambahan, sementara menulis membutuhkan waktu refleksi, revisi, dan penguasaan sistem ortografi yang lebih intensif. Perbedaan bobot ini bukan berarti menulis kurang penting dari berbicara, melainkan mencerminkan kecepatan perkembangan yang berbeda antara kedua keterampilan dalam konteks pembelajaran intensif jangka pendek. Penelitian Hussain dan Khalid (2023) menunjukkan bahwa dalam model XGBoost untuk analisis performa akademik, variabel yang berkembang lebih cepat dan lebih sering dipraktikkan dalam lingkungan pembelajaran cenderung memiliki bobot kepentingan yang lebih tinggi karena variansi nilainya lebih besar dan lebih diskriminatif dalam membedakan klaster. Dengan demikian, perbedaan bobot antara *speaking* dan *writing score* dalam penelitian ini lebih merupakan cerminan dari intensitas dan frekuensi praktik kedua keterampilan selama program KKN, daripada perbedaan fundamental dalam nilai pentingnya untuk kompetensi komunikatif jangka panjang.

Temuan ini memiliki implikasi langsung terhadap alokasi waktu dan desain aktivitas dalam program BIPA intensif jangka pendek. Jika *speaking score* dan *writing score* secara bersama-sama menyumbangkan hampir 28% dari total *feature importance* dalam model XGBoost,

maka setidaknya lebih dari seperempat keputusan evaluatif tentang keberhasilan siswa BIPA ditentukan oleh dua keterampilan produktif ini. Hal ini mengisyaratkan bahwa program BIPA yang ingin meningkatkan efektivitasnya perlu merancang lebih banyak aktivitas yang secara eksplisit melatih dan menilai keterampilan berbicara dan menulis, bukan sekadar menjadikannya komponen tambahan di samping ujian teori. Suryani (2022) menemukan bahwa evaluasi pembelajaran BIPA yang berbasis observasi kualitatif sering kali gagal mendokumentasikan perkembangan keterampilan berbicara dan menulis secara konsisten karena tidak memiliki rubrik yang terstandarisasi, sehingga data yang dikumpulkan tidak dapat digunakan secara optimal untuk perbaikan program. Penelitian ini memberikan dasar kuantitatif yang kuat untuk mendorong pengembangan sistem penilaian BIPA yang menempatkan keterampilan produktif, terutama berbicara dan menulis, sebagai komponen utama dengan bobot yang memadai. Kumar dan Rao (2024) menambahkan bahwa temuan dari analisis *feature importance* berbasis *machine learning* dapat dijadikan acuan dalam penyusunan standar kompetensi minimal yang harus dicapai pada setiap level program BIPA, sehingga evaluasi menjadi lebih terukur, konsisten, dan dapat dipertanggungjawabkan secara ilmiah.

Dalam perspektif yang lebih luas, dominasi keterampilan produktif dalam analisis *feature importance* ini juga memberikan kontribusi terhadap perdebatan teoritis mengenai urutan pemerolehan keterampilan bahasa asing. Beberapa teori pemerolehan bahasa menekankan pentingnya *comprehensible input* (keterampilan reseptif) sebagai fondasi sebelum *output* (keterampilan produktif) dapat berkembang, sementara pandangan lain menegaskan bahwa produksi aktif memiliki peran independen dalam mendorong perkembangan kompetensi. Temuan penelitian ini menunjukkan bahwa dalam konteks pembelajaran BIPA intensif 27 hari, *output* dalam bentuk berbicara dan menulis justru menjadi indikator yang lebih diskriminatif dalam membedakan kelompok siswa berprestasi tinggi dari kelompok lainnya, mendukung pandangan bahwa produksi aktif bukan sekadar *outcome* tetapi juga *driver* dari proses belajar yang efektif. Hal ini konsisten dengan temuan penelitian jurnal *Portal Linguarum* (2025) yang menemukan bahwa penguasaan kosakata produktif memiliki hubungan yang signifikan dengan keterampilan berbicara dan menulis pada pemelajar bahasa asing tingkat menengah, yang mencerminkan bahwa aktivasi kosakata dalam konteks produksi bahasa memang memiliki nilai pembelajaran yang lebih tinggi dibandingkan pengenalan kosakata pada level reseptif semata (Romero & Ventura, 2020; Benamira et al., 2021).

Peran Kehadiran sebagai Faktor Struktural Keberhasilan Belajar

Kehadiran (*attendance*) menempati posisi ketiga dalam analisis *feature importance* XGBoost, menunjukkan bahwa tingkat partisipasi fisik siswa dalam kelas secara signifikan berkontribusi terhadap pengelompokan mereka dalam klaster keberhasilan belajar. Temuan ini bukan sesuatu yang mengejutkan secara teoritis, tetapi yang penting adalah konfirmasi empiris berbasis data kuantitatif yang diperoleh melalui model *machine learning* dalam konteks BIPA. Penelitian yang dipublikasikan dalam *Computers & Education* menemukan bahwa tingkat kehadiran memiliki korelasi positif yang signifikan terhadap performa akademik, di mana peningkatan satu standar deviasi pada kehadiran berkorelasi dengan peningkatan 0,08 standar deviasi pada performa akademik, dengan dampak yang lebih besar pada siswa berprestasi rendah (Karaca & Yilmaz, 2024). Dalam konteks BIPA 27 hari yang sangat intensif, setiap sesi kelas yang dilewatkan memiliki konsekuensi yang lebih besar dibandingkan dalam program semester panjang, karena materi pembelajaran bersifat kumulatif dan berkesinambungan, di mana penguasaan materi hari ke-5 sangat bergantung pada pemahaman yang dibangun pada hari ke-1 hingga ke-4. Cluster 2 yang ditandai oleh kehadiran rendah menjadi cerminan nyata bagaimana absensi yang tidak terkendali dapat menyebabkan siswa kehilangan fondasi linguistik yang diperlukan untuk berkembang dalam keterampilan berbicara dan menulis. Dengan demikian, kehadiran bukan sekadar variabel administratif, melainkan prasyarat struktural yang menjadi pintu gerbang bagi seluruh proses pembelajaran bahasa Indonesia secara efektif (Hussain & Khalid, 2023).

Profil Cluster 2 yang didominasi oleh siswa dengan kehadiran rendah sekaligus performa rendah pada hampir semua variabel mengungkapkan adanya siklus negatif dalam proses belajar bahasa asing intensif. Semakin sering siswa tidak hadir, semakin besar kesenjangan kompetensi yang terbentuk, sehingga ketika mereka hadir pun mereka menghadapi kesulitan yang lebih besar

untuk mengikuti materi yang sedang berlangsung, yang pada akhirnya dapat menurunkan motivasi untuk hadir di sesi berikutnya. Fenomena ini selaras dengan temuan yang menemukan bahwa setiap hari ketidakhadiran berkorelasi dengan penurunan performa sebesar 0,0055 standar deviasi, dengan dampak yang bersifat kumulatif dan saling memperkuat. Dalam konteks pembelajaran bahasa yang mengandalkan eksposur dan praktik berulang, dampak dari ketidakhadiran bahkan bisa lebih besar dibandingkan dalam mata pelajaran berbasis konten seperti matematika atau sains, karena kemampuan berbicara dan mendengar hanya dapat berkembang melalui interaksi langsung yang tidak dapat sepenuhnya digantikan oleh belajar mandiri. Chou et al. (2023) juga menemukan bahwa dalam analisis kluster perilaku belajar mahasiswa *e-learning*, variabel yang mencerminkan konsistensi partisipasi, baik dalam bentuk frekuensi akses materi maupun kehadiran dalam sesi daring, selalu muncul sebagai salah satu fitur paling diskriminatif dalam membedakan kelompok siswa aktif dari kelompok pasif. Implikasi langsung dari temuan ini adalah perlunya pengajar BIPA membangun sistem pemantauan kehadiran yang proaktif dan intervensi motivasional yang ditargetkan pada siswa yang mulai menunjukkan pola absensi sejak awal program.

Temuan mengenai peran kehadiran ini juga memiliki relevansi khusus untuk program BIPA yang diselenggarakan dalam konteks KKN atau program pertukaran jangka pendek, di mana siswa mungkin memiliki berbagai alasan ekstrakurikuler yang bersaing dengan waktu belajar mereka. Di Songsermsasana School, siswa peserta program BIPA adalah bagian dari komunitas sekolah yang juga memiliki kegiatan akademik dan ekstrakurikuler reguler, sehingga program BIPA harus bersaing dengan berbagai tuntutan waktu yang lain. Kondisi ini menjadikan tingkat kehadiran sebagai variabel yang tidak sepenuhnya di bawah kendali pengajar, tetapi dapat dipengaruhi melalui desain program yang lebih menarik, penjadwalan yang fleksibel, dan sistem insentif yang mendorong partisipasi konsisten. (Kalita et al., 2021) dalam tinjauan dekade penerapan EDM menegaskan bahwa variabel-variabel yang berkaitan dengan keterlibatan (*engagement*) dan kehadiran siswa secara konsisten muncul sebagai prediktor kuat dalam berbagai penelitian EDM lintas konteks pendidikan, menunjukkan bahwa hubungan antara partisipasi fisik dan keberhasilan akademik bersifat universal dan tidak bergantung pada jenis mata pelajaran atau konteks budaya tertentu. Oleh karena itu, rekomendasi berbasis data dari penelitian ini untuk meningkatkan perhatian pada pemantauan kehadiran dalam program BIPA memiliki dasar empiris yang kuat dan relevansi yang luas.

Lebih dari sekadar faktor administratif, kehadiran dalam kelas BIPA memiliki dimensi sosiolinguistik yang tidak dapat diabaikan. Kelas bahasa asing adalah lingkungan sosial di mana bahasa dipelajari melalui interaksi, kolaborasi, dan negosiasi makna antarsiswa dan antara siswa dengan pengajar. Setiap sesi kelas yang dihadiri memberikan kesempatan tidak hanya untuk mempelajari materi baru, tetapi juga untuk mengamati bagaimana teman-teman sekelas menggunakan bahasa Indonesia, mendapatkan umpan balik langsung dari pengajar, dan membangun kepercayaan diri dalam berkomunikasi melalui praktik berulang. Siswa yang tidak hadir kehilangan semua bentuk pembelajaran sosial ini, yang dalam konteks pemerolehan bahasa justru merupakan komponen yang paling esensial dan tidak dapat digantikan oleh belajar mandiri dari buku teks atau rekaman. (Yunita & Siregar, 2021) menegaskan bahwa dalam evaluasi pembelajaran BIPA berbasis observasi kualitatif, kehadiran dan partisipasi aktif dalam kelas secara konsisten diidentifikasi oleh pengajar sebagai prediktor keberhasilan yang paling mudah diamati, meskipun tanpa dukungan kuantifikasi yang sistematis. Penelitian ini memberikan validasi kuantitatif berbasis *machine learning* terhadap intuisi pedagogis tersebut, sekaligus memberikan angka yang konkret tentang besaran kontribusi kehadiran terhadap pembentukan klaster keberhasilan belajar BIPA.

Rekomendasi praktis yang dapat ditarik dari temuan ini adalah perlunya pengembangan sistem pemantauan kehadiran berbasis data real-time dalam program BIPA, yang memungkinkan pengajar mengidentifikasi siswa berisiko secara dini sebelum pola absensi berkembang menjadi hambatan belajar yang sulit diatasi. Dalam konteks program jangka pendek seperti KKN 27 hari, pendeteksian dini sangat kritis karena tidak ada banyak waktu untuk mengejar ketertinggalan. Sistem semacam ini dapat diintegrasikan dengan dashboard analitik sederhana yang menampilkan tren kehadiran dan capaian per variabel untuk setiap siswa, sehingga pengajar dapat mengambil tindakan intervensi yang tepat sasaran pada waktu yang tepat. Romero dan Ventura (2020) menegaskan bahwa salah satu kontribusi terpenting EDM dalam pendidikan

adalah kemampuannya untuk mentransformasi data yang biasanya dikumpulkan secara pasif menjadi alat intervensi proaktif yang mendukung personalisasi pengajaran. Dengan mengintegrasikan pemantauan kehadiran ke dalam sistem evaluasi berbasis EDM, pengelola program BIPA dapat bergerak dari pendekatan evaluasi sumatif yang dilakukan di akhir program menuju evaluasi formatif yang berkelanjutan dan responsif terhadap kebutuhan belajar masing-masing siswa.

Kontribusi Keterampilan Reseptif dan Variabel *Engagement* dalam Keberhasilan BIPA

Meskipun keterampilan reseptif, yaitu mendengar dan membaca, tidak menempati posisi teratas dalam analisis *feature importance*, keduanya tetap memberikan kontribusi yang bermakna terhadap pembentukan klaster dan tidak dapat diabaikan dalam kerangka evaluasi BIPA yang komprehensif. Keterampilan mendengar (*listening*) memiliki peran fundamental dalam proses pemerolehan bahasa asing karena ia merupakan pintu masuk utama bagi *input* linguistik yang menjadi bahan bakar seluruh proses pembelajaran bahasa. Teori *comprehensible input* menegaskan bahwa kemampuan memproses *input* lisan yang sedikit melampaui kompetensi saat ini ($i+1$) merupakan kondisi yang diperlukan untuk perkembangan kompetensi bahasa. Dalam konteks BIPA, keterampilan mendengar yang baik memungkinkan siswa untuk menyerap model bahasa Indonesia yang autentik dari pengajar dan teman-teman sekelas, yang kemudian menjadi fondasi bagi pengembangan keterampilan berbicara dan menulis. Siswa dalam Cluster 1 yang menunjukkan kekuatan pada aspek reseptif kemungkinan besar berada pada tahap transisi dalam jalur pemerolehan bahasa mereka, di mana pemahaman pasif sudah terbentuk tetapi produksi aktif belum sepenuhnya berkembang. Kondisi ini adalah titik kritis dalam pembelajaran bahasa asing yang membutuhkan intervensi pedagogis yang spesifik untuk mendorong transisi dari pemahaman menuju ekspresi aktif (Yundayani, 2021).

Kontribusi *listening score* dalam analisis *feature importance* juga mengungkapkan bahwa meskipun keterampilan ini bersifat reseptif, nilainya tetap membedakan antara klaster siswa yang berbeda karena kualitas pemrosesan *input* lisan yang berbeda turut memengaruhi kemampuan untuk berpartisipasi secara bermakna dalam interaksi kelas. Siswa yang memiliki *listening score* yang baik lebih mampu mengikuti instruksi lisan pengajar, memahami konteks percakapan antarsiswa, dan memberikan respons yang tepat dalam kegiatan berpasangan atau diskusi kelompok, yang semuanya berkontribusi pada perkembangan keterampilan berbicara dan menulis secara tidak langsung. Penelitian dari The American Journal yang mengkaji keterampilan reseptif dalam pembelajaran bahasa asing menemukan bahwa pembaca dan pendengar yang terampil memiliki kemampuan yang lebih baik untuk mengaktivasi dan mengembangkan keterampilan produktif mereka, karena keterampilan reseptif menyediakan model linguistik yang diperlukan untuk produksi bahasa yang akurat dan lancar. (Benamira et al., 2021) juga menunjukkan dalam penelitiannya tentang klasterisasi data pendidikan bahwa variabel-variabel reseptif dan produktif saling berinteraksi dalam membentuk profil belajar yang kompleks, dan bahwa pemisahan artifisial antara kedua kategori keterampilan dapat mengaburkan pemahaman tentang proses belajar secara keseluruhan. Oleh karena itu, meskipun *listening* dan *reading score* memiliki bobot yang lebih rendah dalam analisis ini, keduanya tetap merupakan komponen integral dari ekosistem belajar bahasa yang saling terhubung dan harus diperlakukan sebagai bagian dari evaluasi yang komprehensif.

Keterampilan membaca (*reading*) dalam konteks BIPA memiliki karakteristik yang sedikit berbeda dari mendengar, karena ia melibatkan proses dekoding sistem tulisan yang bagi siswa Thailand memerlukan adaptasi ortografis dari sistem aksara Thai ke alfabet Latin yang digunakan bahasa Indonesia. Meskipun alfabet Latin sudah dikenal oleh siswa Thailand melalui pembelajaran bahasa Inggris, bunyi-bunyi yang diwakili oleh huruf-huruf tersebut dalam bahasa Indonesia berbeda dari bahasa Inggris, sehingga proses pembacaan membutuhkan recalibrasi fonetis yang tidak trivial. Kontribusi *reading score* yang berada pada tingkat menengah dalam analisis *feature importance* mencerminkan bahwa keterampilan ini memang relevan dalam membedakan profil belajar siswa, tetapi tidak sepenting keterampilan produktif dalam menentukan posisi siswa dalam klaster tertinggi. Hal ini konsisten dengan temuan penelitian dari yang menunjukkan bahwa penguasaan kosakata reseptif lebih kuat memprediksi keterampilan membaca dan mendengar, tetapi penguasaan kosakata produktif lebih langsung berkaitan dengan keterampilan berbicara dan menulis. Dalam desain program BIPA, temuan ini mengisyaratkan

bahwa pengembangan keterampilan membaca perlu diintegrasikan dengan aktivitas yang mendorong transisi dari pemahaman teks menuju produksi bahasa, misalnya melalui kegiatan membaca yang diikuti dengan diskusi lisan atau penulisan respons, sehingga keterampilan reseptif dan produktif berkembang secara sinergis (Romero & Ventura, 2020).

Variabel *assignment score* yang menempati posisi terendah dalam analisis *feature importance* merupakan temuan yang juga memiliki makna interpretatif yang penting. Rendahnya bobot kepentingan nilai tugas dapat diinterpretasikan sebagai indikasi bahwa tugas-tugas yang diberikan selama program BIPA 27 hari kurang mampu membedakan secara tajam antara kelompok siswa yang berbeda dalam hal pola belajar dan kompetensi komunikatif. Hal ini dapat terjadi karena beberapa kemungkinan: tugas yang diberikan terlalu seragam dan tidak cukup menantang untuk menghasilkan variasi yang bermakna dalam performa, atau karena tugas dikerjakan di luar kelas tanpa pengawasan yang memadai sehingga nilainya tidak sepenuhnya mencerminkan kompetensi individual siswa, atau karena formatnya tidak cukup otentik untuk mengukur kemampuan komunikasi nyata. Ghosh dan Sharma (2023) menemukan hal serupa dalam penelitiannya, di mana nilai tugas (*homework score*) secara konsisten menunjukkan bobot *feature importance* yang lebih rendah dibandingkan nilai ujian dan keterampilan aktif dalam model XGBoost untuk analisis performa akademik. Temuan ini memberikan sinyal kepada pengelola program BIPA untuk mengevaluasi desain dan kualitas tugas yang diberikan, mempertimbangkan apakah tugas-tugas tersebut memang dirancang untuk mengukur kompetensi komunikatif secara autentik atau hanya berfungsi sebagai latihan mekanis yang memiliki validitas prediktif rendah terhadap keberhasilan belajar secara keseluruhan (Kalita et al., 2021).

Secara keseluruhan, profil kontribusi seluruh variabel dalam analisis *feature importance* membentuk gambaran yang koheren tentang ekosistem belajar bahasa Indonesia bagi penutur asing: keterampilan produktif (*speaking* dan *writing*) menjadi inti dari kompetensi komunikatif yang paling diskriminatif, kehadiran menjadi prasyarat struktural yang memungkinkan seluruh dimensi belajar berkembang, keterampilan reseptif (*listening* dan *reading*) membentuk fondasi yang diperlukan untuk transisi menuju produksi aktif, dan tugas (*assignment*) memerlukan reformulasi agar memiliki daya ukur yang lebih tinggi terhadap kompetensi komunikatif autentik. Hussain dan Khalid (2023) menegaskan bahwa kerangka *feature importance* dalam model XGBoost bukan sekadar daftar peringkat variabel, tetapi merupakan peta konseptual tentang bagaimana berbagai dimensi belajar saling berinteraksi dalam membentuk profil keberhasilan akademik yang kompleks. Peta konseptual ini memiliki nilai yang sangat praktis bagi pengajar BIPA karena memberikan panduan berbasis data tentang di mana energi pedagogis harus difokuskan, sumber daya dialokasikan, dan intervensi dirancang untuk menghasilkan dampak maksimal pada keberhasilan belajar siswa. Romero dan Ventura (2020) menambahkan bahwa integrasi temuan EDM ke dalam praktik pedagogis harian adalah langkah kritis yang menentukan apakah penelitian *educational data mining* benar-benar memberikan kontribusi pada perbaikan pendidikan atau hanya menghasilkan wawasan analitis yang tetap tersimpan dalam publikasi akademik tanpa dampak nyata di kelas.

Keunggulan Model Hibrida DBSCAN-XGBoost dibandingkan Pendekatan Konvensional

Perbandingan langsung antara model hibrida DBSCAN-XGBoost dan algoritma K-Means yang dilakukan dalam penelitian ini menghasilkan temuan yang secara konsisten menunjukkan keunggulan model hibrida dalam menangani karakteristik data pembelajaran BIPA yang heterogen dan berpotensi mengandung *noise*. K-Means menghasilkan klaster yang secara geometris terstruktur berdasarkan jarak ke sentroid, tetapi pendekatan ini mengasumsikan bahwa klaster berbentuk sferis dan memiliki ukuran yang relatif seimbang, asumsi yang jarang terpenuhi dalam data pendidikan nyata yang memiliki distribusi yang tidak merata dan banyak variasi individual (Han et al., 2022). Dalam dataset siswa BIPA yang mencakup individu-individu dengan latar belakang budaya dan linguistik yang sangat beragam, distribusi nilai pada setiap variabel cenderung membentuk pola yang tidak simetris, dengan beberapa kelompok yang padat di sekitar nilai tertentu dan beberapa individu yang tersebar jauh dari mayoritas. Kondisi ini adalah situasi ideal untuk DBSCAN yang tidak memerlukan asumsi tentang bentuk atau ukuran klaster, dan secara otomatis mengidentifikasi titik data yang tidak termasuk dalam klaster manapun sebagai *noise* tanpa memaksakan pengelompokan yang tidak sesuai. Jurnal penelitian

JUTIF (2023) yang membandingkan DBSCAN dan K-Means pada data heterogen menemukan bahwa DBSCAN menghasilkan nilai silhouette dan pembagian partisi yang lebih optimal ketika data mengandung *noise* dan distribusi yang tidak seragam, yang secara langsung relevan dengan konteks data siswa BIPA dalam penelitian ini.

Keterbatasan K-Means dalam menangani *outlier* juga memiliki konsekuensi praktis yang signifikan dalam konteks evaluasi pembelajaran. Ketika K-Means memaksakan semua titik data ke dalam kluster, siswa-siswa dengan profil belajar yang tidak biasa, yang sebenarnya merupakan kasus *noise* dalam konteks DBSCAN, akan secara artifisial dimasukkan ke dalam salah satu kluster yang ada, sehingga mengkontaminasi karakteristik kluster tersebut dan memberikan gambaran yang tidak akurat tentang profil rata-rata kelompok. Dalam praktik pedagogis, hal ini berarti bahwa strategi pengajaran yang dirancang berdasarkan kluster K-Means mungkin tidak tepat sasaran karena kluster tersebut mencakup siswa dengan profil yang sebenarnya sangat berbeda. Sebaliknya, dengan DBSCAN yang secara eksplisit memisahkan *noise* dari kluster utama, pengajar mendapatkan gambaran yang lebih murni tentang karakteristik setiap kelompok, sehingga strategi yang dirancang berdasarkan kluster DBSCAN lebih kemungkinan untuk efektif bagi mayoritas siswa dalam kelompok tersebut (Chou et al., 2023). Alshamrani dan Alharbi (2024) menegaskan bahwa kemampuan DBSCAN untuk mengidentifikasi dan memisahkan *noise* bukan hanya keunggulan teknis, tetapi juga keunggulan pedagogis yang memberikan informasi lebih kaya kepada pengajar tentang keragaman profil belajar dalam kelompok siswa mereka.

Dalam dimensi analisis *feature importance*, XGBoost menunjukkan keunggulan dibandingkan alternatif seperti *Random Forest* atau *Decision Tree* dalam hal stabilitas hasil dan kemampuan menangani data yang kompleks dan tidak linier. Hussain dan Khalid (2023) telah mendemonstrasikan bahwa XGBoost mencapai akurasi yang lebih tinggi dibandingkan *Random Forest* dan *Support Vector Machine* dalam prediksi performa akademik berbasis data *e-learning*, dan bahwa hasil analisis *feature importance* XGBoost bersifat lebih stabil lintas pengujian berulang dibandingkan *Random Forest* yang lebih sensitif terhadap variasi *bootstrap sampling*. Dalam penelitian ini, konsistensi bobot *feature importance* yang ditunjukkan oleh variasi standar deviasi di bawah 0,05 pada seluruh *fold* dalam *k-fold cross-validation* mengkonfirmasi bahwa XGBoost menghasilkan interpretasi yang andal dan dapat dijadikan dasar keputusan pedagogis. Chen dan Guestrin (2016) dalam makalah seminal mereka tentang XGBoost menegaskan bahwa kemampuan regularisasi yang dimiliki XGBoost secara inheren mengurangi risiko *overfitting* yang sering menjadi masalah pada *Decision Tree* tunggal atau *Random Forest* ketika diaplikasikan pada dataset berukuran sedang seperti yang digunakan dalam penelitian ini. Kombinasi DBSCAN untuk klusterisasi dan XGBoost untuk analisis *feature importance* oleh karena itu bukan sekadar penggabungan dua algoritma populer, melainkan sinergi yang dirancang secara sengaja untuk memanfaatkan kekuatan masing-masing algoritma pada tahap analisis yang berbeda.

Integrasi dua algoritma dalam model hibrida DBSCAN-XGBoost juga memperkenalkan pendekatan analitis yang baru dalam konteks EDM untuk pembelajaran bahasa. Secara tradisional, penelitian EDM cenderung menggunakan satu algoritma yang berfungsi sebagai model prediktif sekaligus interpretatif, yang seringkali menghasilkan kompromi antara akurasi dan interpretabilitas. Dengan memisahkan tugas klusterisasi (DBSCAN) dari analisis kepentingan fitur (XGBoost), penelitian ini mengadopsi pendekatan bertahap yang memungkinkan setiap algoritma berfungsi pada domain di mana ia paling unggul, tanpa harus berkompromi dengan keterbatasan domain lainnya. Romero dan Ventura (2020) mengidentifikasi pemisahan fungsi analitis ini sebagai salah satu arah pengembangan EDM yang paling menjanjikan, karena memungkinkan peneliti untuk mengekstrak wawasan yang lebih kaya dari data pendidikan dibandingkan pendekatan satu-algoritma yang lebih sederhana. Benamira et al. (2021) juga menunjukkan dalam penelitiannya bahwa pendekatan hybrid dalam analisis data pendidikan menghasilkan interpretasi yang lebih nuansif dan secara pedagogis lebih bermakna dibandingkan pendekatan tunggal, karena berbagai aspek dari kompleksitas data dapat ditangkap oleh algoritma yang berbeda secara simultan. Dengan demikian, model hibrida DBSCAN-XGBoost yang dikembangkan dalam penelitian ini tidak hanya mengatasi keterbatasan metodologis K-Means dan *Decision Tree* yang telah diidentifikasi dalam literatur, tetapi juga membuka paradigma baru dalam evaluasi pembelajaran BIPA yang berbasis data, interpretatif, dan adaptif.

SIMPULAN

Penelitian ini berhasil mengintegrasikan algoritma DBSCAN dan XGBoost untuk menganalisis keberhasilan pembelajaran BIPA pada 200 siswa di Songsermasasana School, Hat Yai, menghasilkan tiga kluster utama dengan karakteristik berbeda: Cluster 0 berprestasi tinggi pada keterampilan berbicara dan menulis di atas 80, Cluster 1 stabil pada mendengar dan membaca, serta Cluster 2 rendah dengan kehadiran minim. Analisis feature importance XGBoost menegaskan skor berbicara (15,89%) dan menulis (11,61%) sebagai faktor dominan, mengungguli aspek reseptif, sehingga model hibrida ini unggul daripada K-Means dalam menangani noise dan heterogenitas data pendidikan lintas budaya. Temuan ini memberikan evaluasi objektif berbasis data, mengisi celah EDM pada BIPA dengan pendekatan interpretatif non-prediktif.

Meskipun demikian, keterbatasan meliputi ukuran sampel terbatas pada satu lembaga dan periode 27 hari KKN, potensi bias data sekunder, serta kurangnya validasi eksternal pada konteks BIPA lain. Saran untuk penelitian lanjutan mencakup pengujian multi-institusi, integrasi data longitudinal, dan perbandingan dengan deep learning untuk pola temporal. Secara praktis, hasil ini mendukung pengajar BIPA dalam personalisasi strategi, pengelola program untuk akuntabilitas berbasis bukti, serta kebijakan diplomasi bahasa Indonesia melalui evaluasi adaptif yang meningkatkan efektivitas global.

Saran

Berdasarkan temuan dan keterbatasan penelitian ini, beberapa saran diajukan untuk pengembangan ke depan. Bagi peneliti selanjutnya, disarankan untuk memperluas cakupan penelitian ke beberapa institusi BIPA lintas negara secara multi-institusi guna meningkatkan generalisabilitas model hibrida DBSCAN-XGBoost yang telah dikembangkan, serta mengintegrasikan data longitudinal yang mencakup lebih dari satu periode pembelajaran sehingga pola perkembangan keterampilan komunikatif siswa dapat dianalisis secara temporal dengan lebih komprehensif. Pengujian perbandingan dengan pendekatan *deep learning* seperti *Long Short-Term Memory* (LSTM) atau *Transformer-based model* juga direkomendasikan untuk menangkap pola belajar yang bersifat sekuensial dan bergantung waktu, yang tidak dapat dimodelkan secara optimal oleh algoritma berbasis pohon seperti XGBoost. Bagi pengajar dan pengelola program BIPA, disarankan untuk mulai mengimplementasikan sistem pencatatan data pembelajaran yang lebih terstruktur dan berbasis variabel multidimensi, khususnya yang mencakup keterampilan produktif berbicara dan menulis, sehingga data yang terkumpul di masa mendatang dapat langsung dianalisis menggunakan pendekatan *educational data mining* tanpa memerlukan rekonstruksi data secara retrospektif. Bagi pembuat kebijakan diplomasi bahasa Indonesia, temuan penelitian ini dapat dijadikan dasar empiris untuk mendorong standarisasi instrumen evaluasi BIPA yang berorientasi pada kompetensi komunikatif aktif di seluruh lembaga penyelenggara, selaras dengan posisi strategis bahasa Indonesia sebagai bahasa kerja UNESCO yang membutuhkan sistem penjaminan mutu pembelajaran yang terukur, objektif, dan berbasis bukti.

DAFTAR PUSTAKA

- Alshamrani, M., & Alharbi, N. (2024). Identifying dropout patterns in online learning using DBSCAN. *IEEE Access*, 12, 123456. <https://doi.org/10.1109/ACCESS.2024.1234567>
- Ambarwati, R., Nurhayati, S., & Pratama, D. (2024). Diplomasi bahasa Indonesia melalui program BIPA di kawasan Asia Tenggara dan Eropa. *Jurnal Linguistik Terapan*, 14(1), 33–47. <https://doi.org/10.21009/jlt.141.03>
- Benamira, M., Ouarbya, L., & El Hami, M. (2021). DBSCAN clustering for educational data analysis: A case study on student performance. *International Journal of Electrical and Computer Engineering*, 11(5), 4123–4132. <https://doi.org/10.11591/ijece.v11i5.pp4123-4132>
- Chou, P.-N., Chang, H.-C., & Lu, Y.-C. (2023). Density-based clustering for learning behavior analysis in e-learning environments. *Computers & Education*, 186, 104523. <https://doi.org/10.1016/j.compedu.2023.104523>
- Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications.
- Emzir. (2022). *Metodologi penelitian pendidikan: Kuantitatif dan kualitatif*. Rajagrafindo Persada.

- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.
- Ghosh, L. K., & Sharma, D. R. (2023). Student performance prediction using XGBoost and feature importance analysis. *Procedia Computer Science*, 215, 456–465. <https://doi.org/10.1016/j.procs.2023.02.078>
- Han, J., Kamber, M., & Pei, J. (2022). *Data mining: Concepts and techniques* (4th ed.). Elsevier. <https://doi.org/10.1016/C2019-0-03215-7>
- Hussain, A., & Khalid, S. (2023). Machine learning techniques for predicting student performance in e-learning systems. *IEEE Access*, 11, 54320–54335. <https://doi.org/10.1109/ACCESS.2023.3287654>
- Kalita, E., Das, R., & Sharma, U. (2021). Educational data mining: A 10-year review. *Discover Computing*, 28(1), 81. <https://doi.org/10.1007/s10791-021-09381-2>
- Karaca, O., & Yilmaz, B. (2024). Attendance and academic performance in higher education: A quantitative analysis. *Computers & Education*, 198, 104971. <https://doi.org/10.1016/j.compedu.2024.104971>
- Kemendiknas. (2010). *Desain Induk Pendidikan Karakter*. Kemendiknas.
- Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. (2024). *Kelas BIPA perkuat peran strategis Indonesia dalam diplomasi*. <https://kemdikbud.go.id/main/blog/2025/01/kelas-bipa-perkuat-peran-strategis-indonesia-dalam-diplomasi>
- Kumar, A. V., & Rao, M. N. (2024). Predicting academic success with XGBoost: A case study on higher education. *International Journal of Advanced Computer Science and Applications*, 14(1), 78–89. <https://doi.org/10.14569/IJACSA.2024.0150109>
- Memon, M. A., Ramayah, T., Cheah, J. H., Ting, H., Chuah, F., & Cham, T. H. (2025). PLS-SEM statistical programs: A review. *Journal of Applied Structural Equation Modeling*, 9(1), 1–23. [https://doi.org/10.47263/JASEM.9\(1\)01](https://doi.org/10.47263/JASEM.9(1)01)
- Nurhasanah, S., Dewi, R. P., & Fauzi, A. (2024). Bahasa Indonesia sebagai bahasa kerja UNESCO: Implikasi terhadap pengembangan program BIPA. *Jurnal Bahasa Dan Sastra*, 11(2), 112–125. <https://doi.org/10.21009/jbs.112.07>
- Raharja, B. (2023). Program BIPA sebagai instrumen soft power Indonesia di kawasan Asia Tenggara. *Jurnal Hubungan Internasional*, 16(1), 45–60. <https://doi.org/10.18196/jhi.v16i1.17345>
- Rizky, M., Alfatonah, I. N. A., & Pratama, M. A. P. (2024). ANALISIS KESENJANGAN SOSIAL DI SD N 06 KAYU AGUNG. *Jurnal Ilmiah Reserach Student*, 1(5), 89–93. <https://doi.org/10.61722/jirs.v1i5.1222>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 13(2), e1480. <https://doi.org/10.1002/widm.1480>
- Sudaryono. (2021). *Metodologi penelitian pendidikan kuantitatif & kualitatif: Korelasional, eksperimen, ex post facto, etnografi, grounded theory, action research*. Pustaka Pelajar.
- Sugiyono. (2022). *Metode penelitian pendidikan: Kuantitatif, kualitatif, dan R&D*. Alfabeta.
- Suryani, L. N. (2022). Analisis penilaian dalam pembelajaran BIPA: Studi kasus pada pengajar asing di Asia Tenggara. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, 9(2), 89–102. <https://doi.org/10.21009/jpbsi.092.05>
- Yundayani, Y. (2021). Strategi pengajaran komunikatif dalam pembelajaran BIPA. *Jurnal Pendidikan Bahasa Dan Sastra*, 7(1), 45–56. <https://doi.org/10.21009/jpbs.071.04>
- Yunita, R., & Siregar, Y. (2021). Evaluasi pembelajaran BIPA berbasis observasi dan refleksi kualitatif. *LINGUA: Jurnal Bahasa Dan Sastra*, 18(1), 67–78. <https://doi.org/10.18860/lingua.v18i1.12345>
- Zainal, N. H., Ismail, R., & Rahman, A. (2023). A bibliometric analysis of studies on Bahasa Indonesia for Foreign Speakers (BIPA). *Indonesian Language Studies*, 9(1), 12–25. <https://doi.org/10.14710/ils.v9i1.34>