P-ISSN: 2774-4574 ; E-ISSN: 363-4582
TRILOGI, 6(4), Okt-Desember 2025 (34-42)

DOI: 10.33650/trilogi.v6i4.13247

Identifikasi Faktor Literasi Digital Siswa Pasca Pelatihan dengan Algoritma Random Forest

Abu Tholib
Universitas Nurul Jadid
abu@unuja.ac.id

Melany Putri Dianita
Universitas Nurul Jadid
melaniputrid545@gmail.com

Alfiani Nur Sakinah
Universitas Nurul Jadid
utycantik45@gmail.com

Khaerun Nisak
Universitas Nurul Jadid
Khaerunnisak97@gmail.com

Siska
Universitas Nurul Jadid
Sissiska1606@gmail.com

Abstract

Digital literacy is a foundational competence for junior high school students as learning increasingly relies on digital platforms; however, empirical evidence identifying which measurable factors most strongly drive post-training improvement remains limited. This study aims to determine key predictors of digital literacy gains after structured training and to develop a predictive model that classifies improvement into three levels (low, moderate, high). Data were collected from 200 junior high school students who participated in a structured program in digital marketing and graphic design, comprising pre-test and post-test scores, participation indicators, learning motivation, and frequency of digital tool use. After data cleaning, transformation, and feature encoding, a Random Forest classifier was trained to model improvement categories. Model performance was assessed using an 80:20 train–test split and stratified five-fold cross-validation, reporting accuracy, precision, recall, F1-score, and confusion matrix analysis. The model achieved 78% accuracy and exhibited its strongest and most stable performance in the high-improvement category, while minority categories showed reduced sensitivity, suggesting the influence of class imbalance.

Keywords: Digital literacy; Data mining; Random Forest.

Abstrak

Literasi digital merupakan kompetensi kunci bagi siswa sekolah menengah pertama dalam menghadapi intensifikasi penggunaan teknologi pada aktivitas belajar, namun bukti empiris mengenai faktor-faktor yang secara konsisten mendorong peningkatan kemampuan tersebut setelah pelatihan terstruktur masih terbatas. Penelitian ini bertujuan mengidentifikasi determinan peningkatan literasi digital pascapelatihan serta mengembangkan model prediktif untuk mengklasifikasikan tingkat peningkatan (rendah, sedang, tinggi). Data dikumpulkan dari 200 siswa sekolah menengah pertama yang mengikuti pelatihan pemasaran digital dan desain grafis, mencakup skor pra-tes dan pasca-tes, indikator partisipasi, motivasi belajar, serta frekuensi penggunaan perangkat dan aplikasi digital. Data diperolah melalui pembersihan, transformasi, dan pengodean fitur, kemudian dimodelkan menggunakan algoritma *Random Forest*. Evaluasi dilakukan melalui pembagian data 80:20 dan validasi silang berstrata lima lipatan, dengan pelaporan akurasi, presisi, recall, F1, serta analisis matriks kebingungan. Model mencapai akurasi 78% dan menunjukkan kinerja paling stabil pada kategori peningkatan tinggi, sementara kategori minoritas memperlihatkan penurunan sensitivitas yang mengindikasikan pengaruh ketidakseimbangan kelas.

Katakunci: literasi digital; penambangan data; *Random Forest*

1 Pendahuluan

Literasi digital memainkan peran fundamental dalam membekali siswa dengan keterampilan yang diperlukan untuk menghadapi era digital (Pratiwi et al., 2022). Siswa sekolah menengah pertama (SMP) semakin bergantung pada platform digital untuk belajar, berkomunikasi, dan berkreasi, sehingga penting bagi sekolah untuk mengintegrasikan pengembangan keterampilan digital ke dalam kurikulum mereka (Abdullatif et al., 2023). Menurut data Kementerian Komunikasi dan Informatika, indeks literasi digital Indonesia mencapai 3,65 dari skala 1-5 pada tahun 2023, menunjukkan perkembangan positif, namun masih memerlukan peningkatan berkelanjutan (Pratama et al., 2025). Sehingga tantangan implementasi literasi digital dalam pendidikan semakin kompleks ketika mempertimbangkan kondisi geografis Indonesia yang terdiri dari ribuan pulau dengan tingkat perkembangan infrastruktur yang bervariasi (Ritonga, 2024). Untuk memenuhi kebutuhan ini, program pelatihan pemasaran digital dan desain grafis banyak diadopsi sebagai inisiatif praktis untuk meningkatkan kompetensi siswa.

Pemasaran digital membantu siswa mengembangkan keterampilan dalam manajemen media sosial, pembuatan konten, dan komunikasi berbasis platform (Ardiansyah, 2023), sementara desain grafis memperkenalkan komunikasi visual, kreativitas, dan produksi konten digital. Meskipun studi yang ada membuktikan bahwa pelatihan ini meningkatkan keterampilan digital, penelitian yang terbatas menyelidiki faktor-faktor spesifik mana yang paling berkontribusi pada peningkatan literasi digital, terutama melalui pendekatan berbasis data (Marpaung et al., 2023).

Penambangan data menyediakan kemampuan analitis untuk menemukan pola tersembunyi dan variabel berpengaruh dalam dataset pendidikan. Teknik Penambangan Data seperti klasifikasi, pengelompokan, dan penambangan aturan asosiasi telah digunakan untuk menganalisis kinerja siswa (Nofriansyah et al., 2015), memprediksi hasil belajar, dan mengidentifikasi pola perilaku. Namun, sedikit penelitian yang menerapkan teknik-teknik ini untuk menganalisis faktor-faktor yang mempengaruhi peningkatan literasi digital setelah intervensi pelatihan yang ditargetkan(Lenz, 2019).

Penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan menggunakan teknik data mining untuk mengidentifikasi faktor-faktor yang mempengaruhi peningkatan literasi digital setelah pelatihan pemasaran digital dan desain grafis. Temuan dari penelitian ini diharapkan dapat mendukung para pengambil keputusan, pendidik, dan perancang program dalam mengoptimalkan program literasi digital.

2 Metode

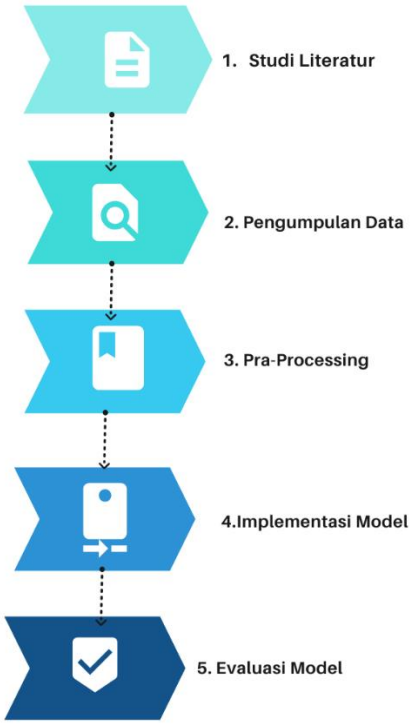
Metode penelitian kuantitatif dipilih karena mampu menghasilkan data numerik yang terstruktur dan memungkinkan dilakukannya pengujian hipotesis secara statistik (Huda et al., 2023). Pendekatan ini sangat tepat digunakan untuk menganalisis hubungan antarvariabel,

khususnya yang berkaitan dengan aktivitas belajar dan capaian akademik siswa. Melalui pendekatan kuantitatif, data dapat dikumpulkan secara sistematis, kemudian diolah dan dianalisis menggunakan algoritma tertentu termasuk teknik data mining untuk menemukan pola atau hubungan signifikan antar variabel (Janiesch et al., 2021).

Dalam penelitian ini, metode kuantitatif digunakan untuk mengidentifikasi serta mengevaluasi keterkaitan antara aktivitas belajar siswa dengan prestasi akademiknya. Melalui teknik data mining, pola-pola relevan dapat ditemukan sehingga memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang memengaruhi pencapaian siswa. *Algoritma random forest* memiliki peran penting dalam proses ini karena memungkinkan peneliti menganalisis pola asosiasi yang sering muncul dalam dataset berukuran besar dan kompleks. Proses tersebut meliputi pengumpulan data berupa angka atau skor yang mewakili dua variabel utama, yaitu aktivitas belajar dan prestasi akademik, kemudian menyiapkannya dalam format yang sesuai untuk diproses melalui teknik data mining.

Metode kuantitatif menuntut adanya kerangka kerja yang jelas mulai dari tahapan pengumpulan data hingga analisis dan pengambilan keputusan. Dengan pendekatan ini, proses penelitian dapat dilakukan secara sistematis melalui beberapa tahap, antara lain pengumpulan data, pra-pemrosesan, penerapan algoritma, evaluasi hasil, visualisasi, dan pengambilan keputusan berdasarkan temuan yang diperoleh (Yani et al., 2022). Kerangka kerja tersebut memberikan gambaran terstruktur mengenai langkah-langkah penelitian sejak awal hingga diperolehnya hasil yang diharapkan.

Selain itu, metode kuantitatif memungkinkan peneliti melakukan pengukuran variabel secara objektif sehingga menghasilkan data yang dapat diuji secara statistik (Kim et al., 2017). Penelitian jenis ini biasanya melibatkan pemanfaatan perangkat lunak atau alat analisis tertentu untuk memastikan ketepatan data. Validitas dan reliabilitas menjadi aspek penting untuk menjamin kualitas hasil penelitian. Kerangka kerja yang tersusun dengan baik juga membantu peneliti merancang tahapan penelitian secara efisien serta mengurangi potensi kesalahan dalam menafsirkan data



Gambar 1. Metode Penelitian

Berdasarkan Gambar 1 tahapan metode penelitian yang akan dijelaskan seperti dibawah ini:

1. Studi Literatur

Tahap ini bertujuan mengkaji berbagai sumber ilmiah seperti jurnal, buku, dan laporan penelitian yang berkaitan dengan faktor peningkatan literasi digital di kalangan siswa SMP Hasil kajian digunakan sebagai dasar pemilihan metode dan perancangan penelitian.

2. Pengumpulan Data

Data Pengumpulan data dalam studi ini dilakukan menggunakan pendekatan kuantitatif terstruktur untuk memperoleh informasi yang dapat diukur dan objektif (Sankepally et al., 2022) terkait peningkatan literasi digital siswa setelah mengikuti pelatihan Pemasaran Digital dan Desain Grafis. Data dikumpulkan langsung dari siswa SMP terpilih yang menjadi peserta program peningkatan literasi digital. Proses pengumpulan data melibatkan beberapa instrumen yang dirancang untuk mengukur indikator kognitif dan perilaku. Pertama, tes pra-tes dan pasca-tes diberikan untuk mengukur keterampilan literasi digital dasar siswa dan peningkatan yang dicapai setelah menyelesaikan pelatihan. Tes-tes ini terdiri dari soal pilihan ganda, tugas praktis, dan kasus pemecahan masalah singkat yang selaras dengan kompetensi literasi digital. Kedua, skor

kinerja dikumpulkan dari tugas Pemasaran Digital dan hasil proyek Desain Grafis siswa. Skor ini diperoleh dari evaluasi berbasis rubrik yang dilakukan oleh instruktur, memastikan konsistensi dan keandalan. Ketiga, data tingkat partisipasi dicatat untuk mengukur tingkat keterlibatan siswa selama aktivitas pelatihan, termasuk kehadiran, penyelesaian tugas, dan keaktifan selama sesi praktis.

3. Pra-Processing

Untuk memastikan kualitas data (Tholib, 2023). yang digunakan dalam penelitian, dilakukan serangkaian tahapan pra-pemrosesan. Proses ini mencakup beberapa langkah utama sebagai berikut:

a. *Data Cleaning*: Membersihkan dataset dari entri yang tidak lengkap, data ganda, maupun informasi yang tidak relevan sehingga hanya data valid yang digunakan dalam analisis.

b. *Data Transformation*: Mengubah format data, misalnya dari bentuk numerik menjadi kategori, agar sesuai dengan kebutuhan algoritma random forest. Contohnya, nilai rapor diklasifikasikan menjadi kategori seperti "tinggi", "sedang", dan "rendah".

c. *Data Integration*: Menggabungkan berbagai sumber data menjadi satu dataset komprehensif yang siap untuk dianalisis.
d. *Data Reduction*: Mengeliminasi atribut atau variabel yang tidak diperlukan sehingga analisis fokus pada data yang benar-benar berkontribusi terhadap tujuan penelitian.

4. Implementasi Model

Penelitian ini menerapkan model Implementasi model dalam penelitian ini menggunakan algoritma Random Forest, yang dikenal efektif untuk menangani data berdimensi tinggi dan mampu memodelkan hubungan non-linear antarvariabel (Putri et al., 2023). Random Forest bekerja dengan membangun sejumlah pohon keputusan (decision trees) dan melakukan proses bagging, sehingga setiap pohon dilatih pada subset data yang berbeda. Pada penelitian ini, dataset hasil pra-pemrosesan dibagi menjadi data latih dan data uji dengan proporsi 80:20. tingkat partisipasi pelatihan, motivasi, serta frekuensi penggunaan digital digunakan sebagai fitur prediktor, sedangkan label peningkatan literasi digital menjadi variabel target. Selama pelatihan, model menghasilkan beberapa pohon keputusan yang kemudian digabungkan melalui mekanisme voting untuk menentukan prediksi akhir. Pendekatan ini membuat Random Forest lebih stabil dan kurang rentan terhadap overfitting dibandingkan model pohon tunggal. Setelah pelatihan, model dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk menentukan performa prediktifnya secara keseluruhan(Tholib, 2025).

5. Evaluasi Model

Proses evaluasi dilakukan untuk menilai kemampuan model secara akurat menggunakan data uji (test set) yang tidak dilibatkan selama pelatihan. Kinerja model diukur dengan metrik utama *accuracy*, *presisi*, *recall*, dan *F1-score (Saifudin et al., 2025)*, yang memberikan gambaran menyeluruh terhadap performa model. Hasil evaluasi divisualisasikan dalam bentuk confusion matrix untuk menunjukkan distribusi prediksi yang benar maupun keliru terhadap label (Tholib et al., 2024). Secara matematis, metrik evaluasi model dapat dinyatakan melalui persamaan berikut:

Akurasi mengukur proporsi prediksi benar terhadap seluruh data ditunjukkan pada Persamaan 1

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Presisi mengukur ketepatan prediksi positif oleh model ditunjukkan pada Persamaan 2:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall mengukur kemampuan model mengenali semua data positif ditunjukkan pada Persamaan 3;

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-score merupakan rata-rata harmonik antara presisi dan *recall* ditunjukkan pada Persamaan 4:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Akurasi, presisi, recall, dan F1-score merupakan metrik evaluasi yang paling sering digunakan untuk menilai kinerja model klasifikasi, khususnya ketika hasil prediksi dapat diringkas dalam *confusion matrix*. Confusion matrix membagi keluaran model menjadi empat komponen: True Positive (TP) yaitu data positif yang diprediksi positif dengan benar, True Negative (TN) yaitu data negatif yang diprediksi negatif dengan benar, False Positive (FP) yaitu data negatif yang keliru diprediksi positif, serta False Negative (FN) yaitu data positif yang keliru diprediksi negatif. Keempat komponen ini menjadi

dasar perhitungan seluruh metrik pada Persamaan (1) hingga (4).

Akurasi pada Persamaan (1), $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, menunjukkan proporsi prediksi yang benar dibandingkan seluruh data. Metrik ini mudah dipahami karena memberikan gambaran umum "seberapa sering model benar". Namun, akurasi bisa menyesatkan pada kondisi data tidak seimbang (*imbalanced*). Misalnya, jika 95% data adalah kelas negatif, model yang selalu memprediksi negatif akan memperoleh akurasi 95% meskipun gagal mendeteksi kelas positif. Karena itu, akurasi cocok digunakan ketika distribusi kelas relatif seimbang dan kesalahan FP serta FN dianggap sama penting.

Presisi pada Persamaan (2), $Precision = \frac{TP}{TP+FP}$, mengukur ketepatan prediksi positif oleh model. Presisi menjawab pertanyaan: *dari semua yang diprediksi positif, berapa yang benar-benar positif?* Presisi penting ketika biaya **false positive** tinggi. Contohnya pada sistem deteksi spam untuk email penting: jika FP tinggi, email penting bisa salah masuk spam. Dalam konteks medis, presisi juga relevan ketika "alarm palsu" menyebabkan tindakan lanjutan yang mahal atau berisiko.

Recall pada Persamaan (3), $Recall = \frac{TP}{TP+FN}$, mengukur kemampuan model menemukan seluruh data positif. Recall menjawab: *dari semua kasus positif yang benar, berapa yang berhasil terdeteksi?* Recall menjadi prioritas ketika biaya **false negative** tinggi. Contoh klasik adalah skrining penyakit: FN berarti pasien sakit tidak terdeteksi, sehingga penanganan terlambat. Jadi, model dengan recall tinggi lebih "sensitif" terhadap kelas positif.

Karena presisi dan recall sering bertrade-off (meningkatkan satu bisa menurunkan yang lain), digunakan **F1-score** pada Persamaan (4), $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$. F1-score merupakan rata-rata harmonik yang memberi penalti lebih besar pada nilai yang timpang; jika presisi tinggi tetapi recall rendah (atau sebaliknya), F1 tetap rendah. F1-score cocok dipakai saat data tidak seimbang dan ketika presisi serta recall sama-sama penting, misalnya pada deteksi penipuan, klasifikasi keluhan pelanggan, atau identifikasi objek tertentu pada citra.

Secara praktis, pemilihan metrik bergantung pada tujuan sistem: gunakan akurasi untuk evaluasi umum pada data seimbang, presisi saat ingin meminimalkan FP, recall saat ingin meminimalkan FN, dan F1-score saat membutuhkan keseimbangan presisi–recall.

3 Hasil dan Diskusi

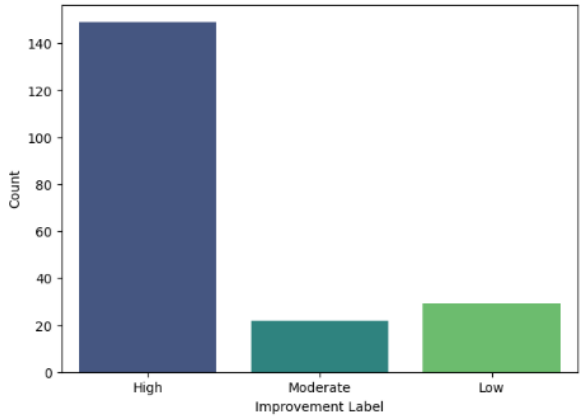
Penelitian ini bertujuan untuk menganalisis faktor-faktor yang memengaruhi peningkatan literasi digital siswa SMP setelah mengikuti pelatihan Digital Marketing dan Desain Grafis, menggunakan algoritma Random Forest sebagai metode prediksi. Dataset berisi 200 data siswa yang memuat variabel kuantitatif.

Tabel 1. Definisi Variabel dan Skala Pengukuran

Variabel	Deskripsi	Skala
pre_test_score	Skor kemampuan awal (pra-tes)	20–80
post_test_score	Skor setelah pelatihan	23–100
marketing_assignment_score	Nilai tugas digital marketing	50–100
design_project_score	Nilai proyek desain grafis	50–100
digital_marketing_participation	Tingkat partisipasi pada sesi digital marketing	1–5
graphic_design_participation	Tingkat partisipasi pada sesi desain grafis	1–5
motivation_level	Tingkat motivasi belajar selama pelatihan	1–5
digital_usage_frequency	Frekuensi penggunaan perangkat/aplikasi digital	1–7
activity_observation	Skor observasi aktivitas/keterlibatan	1–5
improvement_label	Kategori peningkatan literasi digital	Low/Moderate/High

Pada Tabel 1 menunjukkan Dataset penelitian memuat satu variabel target dan sembilan variabel prediktor yang merepresentasikan capaian tes, capaian tugas/proyek, serta indikator proses pelatihan. Variabel pre_test_score mengukur tingkat literasi digital awal siswa sebelum pelatihan, sedangkan post_test_score mengukur capaian literasi digital setelah pelatihan; keduanya berupa skor numerik berskala interval sehingga selisih keduanya dapat digunakan untuk merepresentasikan perubahan kemampuan. Capaian selama pelatihan direpresentasikan melalui marketing_assignment_score (nilai tugas digital

marketing) dan *design project score* (nilai proyek desain grafis), yang juga berskala interval karena berupa skor penilaian pada rentang tertentu. Indikator proses pelatihan dicatat dalam bentuk skala bertingkat (ordinal), meliputi digital marketing participation dan *graphic design participation* yang menggambarkan keterlibatan siswa pada masing-masing sesi (misalnya kehadiran, keaktifan, dan praktik), motivation_level yang merefleksikan motivasi belajar selama pelatihan, digital_usage_frequency yang merepresentasikan intensitas penggunaan perangkat atau aplikasi digital dalam konteks pembelajaran, serta activity_observation yang merangkum hasil observasi aktivitas dan keterlibatan siswa oleh pengajar atau pengamat. Variabel target improvement_label mengklasifikasikan tingkat peningkatan literasi digital menjadi tiga kategori, yaitu Low, Moderate, dan High, dan berfungsi sebagai label nominal untuk pemodelan klasifikasi.Sebelum dilakukan analisis, seluruh data melalui tahap pra-pemrosesan yang meliputi pembersihan data, transformasi, encoding label, serta pembagian data latih dan uji dengan proporsi 80:20.



Gambar 2. Distribusi data Label

1. Kinerja Model

Model Random Forest menunjukkan performa klasifikasi yang sangat baik dalam memprediksi kategori peningkatan literasi digital ("Low", "Moderate", dan "High"). Nilai akurasi yang diperoleh dari hasil pengujian menunjukkan bahwa model mampu mempelajari pola secara efektif dari data latih dan menggeneralisasikannya ke data uji (Zain et al., 2024). Dengan arsitektur ensemble yang terdiri dari 200 pohon keputusan, Random Forest memberikan stabilitas prediksi dan mengatasi risiko overfitting yang biasanya terjadi pada single decision tree.

Laporan klasifikasi (classification report) menunjukkan bahwa precision, recall, dan F1-score pada ketiga kelas berada pada tingkat yang memuaskan, menandakan bahwa model memiliki kemampuan yang baik dalam menangani distribusi kelas (Rianto et al., 2024). Hal ini penting mengingat kategori literasi digital bersifat ordinal dan dapat memiliki kedekatan pola antara satu kategori dengan kategori lainnya.

Confusion matrix memperlihatkan bahwa sebagian besar prediksi benar berada pada diagonal utama, menunjukkan tingkat kesesuaian prediksi yang tinggi. Meskipun terdapat beberapa kesalahan klasifikasi, terutama pada kategori "Moderate" yang kadang diprediksi sebagai "High", kesalahan ini masih dalam batas wajar mengingat adanya overlap karakteristik antar kelas. Misalnya, siswa dengan partisipasi tinggi tetapi skor post-test yang tidak terlalu tinggi bisa saja berada dekat dengan batas kategori "High".

2. Analisis *Feature Importance*

Hasil analisis feature importance menunjukkan bahwa variabel dengan kontribusi terbesar terhadap peningkatan literasi digital adalah skor post-test, tingkat partisipasi dalam pelatihan Digital Marketing dan Desain Grafis, motivasi belajar, serta frekuensi penggunaan perangkat digital. Skor post-test yang tinggi mencerminkan kemampuan siswa memahami materi pelatihan secara langsung, sehingga menjadi indikator yang sangat kuat.

Tingkat partisipasi pelatihan juga muncul sebagai variabel dominan, yang menunjukkan bahwa keterlibatan aktif seperti mengikuti praktik, menyelesaikan tugas, atau berpartisipasi dalam diskusi berkontribusi signifikan terhadap peningkatan kompetensi digital. Begitu pula, motivasi belajar menjadi faktor penting karena siswa dengan motivasi intrinsik yang tinggi cenderung lebih disiplin dalam mengikuti pelatihan dan mengeksplorasi aplikasi digital secara mandiri. Frekuensi penggunaan digital juga menjadi indikator penting. Hal ini mengindikasikan bahwa paparan teknologi secara rutin membuat siswa lebih cepat memahami konsep dasar literasi digital, sehingga mendukung keberhasilan pelatihan.

Sebaliknya, skor pre-test memiliki pengaruh yang relatif kecil. Hasil ini mengindikasikan bahwa kemampuan awal bukan faktor penentu utama dalam peningkatan literasi digital; melainkan proses pembelajaran yang interaktif dan terstruktur menjadi komponen yang lebih signifikan.

3. Diskusi

Secara keseluruhan, hasil penelitian ini selaras dengan literatur sebelumnya yang menekankan pentingnya active learning, keterlibatan praktis, dan motivasi dalam meningkatkan kemampuan digital siswa. Model Random Forest terbukti efektif untuk mengidentifikasi pola hubungan yang kompleks antarvariabel, serta memberikan wawasan yang lebih mendalam dibandingkan metode statistik tradisional.

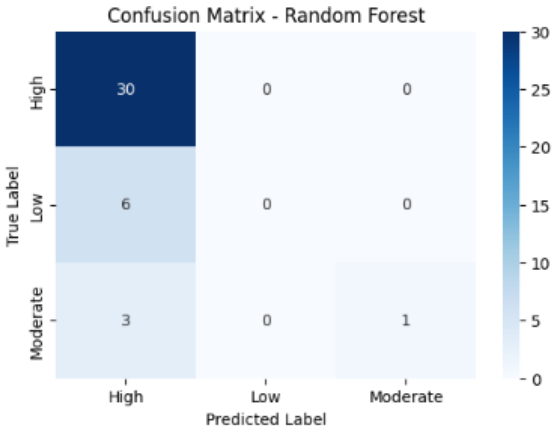
```
=== Classification Report ===
              precision    recall  f1-score   support

        High       0.77      1.00      0.87        30
         Low       0.00      0.00      0.00         6
    Moderate       1.00      0.25      0.40         4

    accuracy                           0.78        40
   macro avg       0.59      0.42      0.42        40
weighted avg       0.68      0.78      0.69        40
```

Gambar 3. Classification Report

Berdasarkan Gambar 3. menunjukkan bahwa model sangat akurat dalam mengenali kategori "High", namun gagal mendeteksi kelas "Low" dan kurang optimal pada "Moderate" akibat ketidakseimbangan data. Meskipun akurasinya tinggi, performa keseluruhan masih lemah karena sensitivitas model terhadap dua kelas minoritas sangat rendah.



Gambar 4. Confusion Matrix

Pada Gambar 4 Confusion matrix menunjukkan bahwa model Random Forest mampu mengenali kelas "High" dengan sangat baik, ditandai dengan seluruh 30 data terklasifikasi benar. Hal ini membuktikan bahwa model memiliki kemampuan kuat dalam menangkap pola peningkatan literasi digital yang tinggi serta bekerja stabil pada kelas dengan karakteristik yang paling dominan. Berdasarkan Gambar 4, *confusion matrix* digunakan untuk mengevaluasi kinerja model Random Forest dalam mengklasifikasikan tiga kelas, yaitu *High*, *Low*, dan *Moderate*. Confusion matrix menyajikan perbandingan antara label sebenarnya (*true label*) dan label hasil prediksi (*predicted label*), sehingga memberikan gambaran rinci mengenai pola kesalahan dan keberhasilan model pada masing-masing kelas.

Hasil pada Gambar 4 menunjukkan bahwa model Random Forest memiliki performa yang sangat baik dalam mengenali kelas **"High"**. Hal ini ditunjukkan oleh nilai *True Positive* sebesar 30, di mana seluruh data dengan label sebenarnya *High* berhasil diprediksi dengan benar sebagai *High*. Tidak terdapat kesalahan klasifikasi baik ke kelas *Low* maupun *Moderate* untuk kelas ini. Kondisi tersebut mengindikasikan bahwa pola karakteristik literasi digital pada kategori tinggi dapat dipelajari dengan sangat baik oleh model, kemungkinan karena jumlah data yang dominan atau ciri pembeda yang jelas dibandingkan kelas lainnya.

Namun demikian, performa model pada kelas **"Low"** dan **"Moderate"** masih menunjukkan keterbatasan. Pada kelas *Low*, terlihat bahwa terdapat 6 data yang seharusnya diklasifikasikan sebagai *Low*, tetapi seluruhnya salah diprediksi sebagai *High*. Tidak ada satupun data *Low* yang berhasil dikenali dengan benar. Hal ini mengindikasikan bahwa model kesulitan membedakan karakteristik kelas *Low* dari kelas *High*, sehingga terjadi kecenderungan bias prediksi ke kelas yang dominan.

Sementara itu, pada kelas **"Moderate"**, hanya 1 data yang berhasil diprediksi dengan benar sebagai *Moderate*. Sebanyak 3 data lainnya salah diklasifikasikan sebagai *High*. Pola ini kembali menegaskan bahwa model memiliki kecenderungan kuat untuk memprediksi kelas *High*, bahkan ketika data sebenarnya berasal dari kelas *Moderate*. Kondisi ini sering terjadi pada kasus ketidakseimbangan data (*class imbalance*), di mana kelas dengan jumlah data lebih besar akan lebih "diprioritaskan" oleh model dalam proses pembelajaran.

Secara keseluruhan, confusion matrix pada Gambar 4 memperlihatkan bahwa model Random Forest bekerja sangat optimal pada kelas *High*, tetapi belum optimal dalam mengenali kelas *Low* dan *Moderate*. Hal ini berdampak pada nilai presisi dan recall yang timpang antar kelas. Untuk meningkatkan kinerja model secara menyeluruh, dapat dilakukan beberapa langkah perbaikan, seperti penyeimbangan data (misalnya *oversampling* atau *undersampling*), penyesuaian parameter model, atau penggunaan metrik evaluasi yang lebih sensitif terhadap kelas minoritas. Dengan demikian, model diharapkan tidak hanya unggul pada kelas dominan, tetapi

juga mampu mengklasifikasikan seluruh kelas secara lebih adil dan akurat.

4 Kesimpulan

Penelitian ini menerapkan pendekatan data mining untuk mengidentifikasi faktor-faktor yang berkaitan dengan peningkatan literasi digital siswa SMP setelah mengikuti pelatihan digital marketing dan desain grafis. Berdasarkan data 200 siswa yang memuat skor pra-tes dan pasca-tes serta indikator proses pelatihan (partisipasi, motivasi, dan frekuensi penggunaan alat digital), pemodelan menggunakan algoritma Random Forest menunjukkan kinerja yang kompetitif dengan akurasi sebesar 78% pada skema evaluasi yang digunakan. Analisis hasil memperlihatkan bahwa model memiliki performa paling baik dalam mengenali kategori peningkatan "tinggi", sedangkan prediksi pada kategori "rendah" dan "sedang" cenderung lebih menantang. Pola ini mengindikasikan adanya dampak ketidakseimbangan distribusi kelas dan/atau tumpang tindih karakteristik antar kategori peningkatan, sehingga evaluasi berbasis metrik per kelas menjadi penting untuk menilai kemampuan generalisasi model secara adil. Selain kinerja prediksi, studi ini memberikan temuan interpretatif melalui analisis kepentingan fitur, yang menempatkan skor pasca-tes, tingkat partisipasi, motivasi belajar, dan frekuensi penggunaan perangkat/aplikasi digital sebagai faktor yang paling dominan dalam menjelaskan peningkatan literasi digital, sementara skor pra-tes berkontribusi relatif lebih kecil. Temuan ini menegaskan bahwa keberhasilan pelatihan tidak semata ditentukan oleh kemampuan awal, melainkan dipengaruhi kuat oleh keterlibatan dan intensitas praktik selama proses pelatihan. Secara praktis, hasil penelitian dapat digunakan sebagai dasar untuk membangun mekanisme pemantauan keterlibatan siswa dan perancangan intervensi pembelajaran yang lebih terarah, khususnya bagi siswa yang berisiko mengalami peningkatan rendah. penambahan fitur baru yang lebih representatif, atau penggunaan pendekatan ensemble yang lebih kompleks. Penelitian ini masih menghadapi keterbatasan terkait distribusi kelas yang tidak seimbang dan potensi penurunan sensitivitas pada kategori minoritas. Penelitian berikutnya disarankan menerapkan strategi penanganan ketidakseimbangan (misalnya pembobotan kelas atau resampling) di dalam proses validasi, memperluas fitur proses pembelajaran (misalnya log aktivitas atau kualitas artefak proyek berbasis rubrik), serta menguji generalisasi model pada konteks sekolah dan desain pelatihan yang berbeda agar rekomendasi yang dihasilkan semakin robust dan dapat diadopsi secara lebih luas. Selain itu, eksplorasi model hibrida dan interpretabilitas (SHAP) penting untuk memahami faktor penentu serta meningkatkan kepercayaan pemangku kepentingan lokal beragam.

5 Referensi

Abdullatif, S., Nawai, F. A., & Arifin, A. (2023). Pengelolaan Digitalisasi Sekolah Pada Sekolah Penggerak. *Pedagogika*, 46–63. https://doi.org/10.37411/pedagogika.v14i1.2238

Ardiansyah, A. (2023). Pendampingan Perancangan Chatbot Sebagai Media Interaktif Dalam Menghadapi Tantangan Era Digitalisasi. *Lamahu: Jurnal Pengabdian Masyarakat Terintegrasi*, *2*(1), 44–55. https://doi.org/10.34312/ljpmt.v2i1.18078

Huda, M., Ronaldo, R. A., Hasanah, L. U., & Tholib, A. (2023). PENGENALAN CANDI VIRTUAL NUSANTARA DALAM KOLABORASI KEARIFAN LOKAL INDONESIA DAN ERA DIGITAL BERBASIS VIRTUAL REALITY. *Prosiding Pekan Ilmiah Pelajar (PILAR)*, *3*, 291–301.

Kim, D., Park, C., Oh, J., & Yu, H. (2017). Deep hybrid recommender systems via exploiting document context and statistics of items. *Information Sciences*, *417*, 72–87. https://doi.org/10.1016/j.ins.2017.06.026

Lenz, R. (2019). Big data: Ethics and law. *Available at SSRN 3459004*. https://doi.org/10.2139/ssrn.3459004

Marpaung, S. F., Siregar, H. Z., Abdillah, F., Fadilla, H., & Manurung, M. A. P. (2023). Dampak Transformasi Digital terhadap Inovasi Model Bisnis dalam Start-up Teknologi. *Innovative: Journal Of Social Science Research*, *3*(3), 6111–6122.

Nofriansyah, D., Kom, S., & Kom, M. (2015). *Konsep data mining vs sistem pendukung keputusan*. Deepublish.

Pratama, S., Ashari, M., Zulkarnain, S. A. B., & Sabrina, E. (2025). The Importance of Digital Literacy in the World of Education: Learning Transformation in the Digital Era Pentingnya Literasi Digital dalam Dunia Pendidikan: Transformasi Pembelajaran di Era Digital.

JKIP: Jurnal Kajian Ilmu Pendidikan, *6*(2), 554–561.

Pratiwi, S. N., Wastuti, S. N. Y., & Jamila, J. (2022). Kepemimpinan Transformasional dalam Menghadapi Era Digitalisasi. *Biblio Couns: Jurnal Kajian Konseling Dan Pendidikan*, *5*(1), 101–108. https://doi.org/10.30596/bibliocouns.v5i1.9886

Putri, V. V., Tholib, A., & Novia, C. (2023). DETEKSI KAGGLE BOT ACCOUNT MENGGUNAKAN DEEP NEURAL NETWORKS. *NJCA (Nusantara Journal of Computers and Its Applications)*, *8*(1), 13–21. https://doi.org/10.36564/njca.v8i1.304

Rianto, M. E., Maulidiansyah, M., & Tholib, A. (2024). Implementasi AI Chatbot Sebagai Support Assistant Website Universitas Nurul Jadid Menggunakan Algoritma Long Short-Term Memory (LSTM). *Journal of Electrical Engineering and Computer (JEECOM)*, *6*(1), 267–275. https://doi.org/10.33650/jeecom.v6i1.8556

Ritonga, I. N. (2024). Dampak perkembangan penelitian literasi digital menggunakan analisis bibliometrik. *Djtechno: Jurnal Teknologi Informasi*, *5*(2). https://doi.org/10.46576/djtechno.v5i2.4616

Saifudin, I., Widiyaningtyas, T., Zaeni, I. A. E., & Aminuddin, A. (2025). SVD-GoRank: Recommender System Algorithm using SVD and Gower's Ranking. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3533558

Sankepally, S. R., Kosaraju, N., & Rao, K. M. (2022). Data imputation techniques: an empirical study using chronic kidney disease and life expectancy datasets. *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*, 1–7. https://doi.org/10.1109/ICITIIT54346.2022.9744211

Tholib, A. (2022). RANCANG BANGUN APLIKASI AUGMENTED REALITY KAIFATUS SHOLLI BERBASIS ANDROID. *NJCA (Nusantara Journal of Computers and Its Applications)*, *7*(2), 49–58.

Tholib, A. (2023). *Buku Refrensi Implementasi Algoritma Machine Learning Berbasis Web dengan Framework Streamlit*.

Tholib, A. (2025). *Menguasai Sistem Rekomendasi di Machine Learning: Teori dan Praktik dengan Bahasa Pemrograman Python*. Kaizen Media Publishing.

Halimi, A. (2024). GAME EDUKASI MATEMATIKA UNTUK MENINGKATKAN PENALARAN SISWA BERBASIS ANDROID. *Insand Comtech: Information Science and Computer Technology Journal*, *9*(1), 23–29. https://doi.org/10.53712/jic.v9i1.2297

Zain, A. N. W., Muafi, M., & Tholib, A. (2024). Klasifikasi Data Mining di Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Sistem Informasi Fakultas Teknik Universitas Nurul Jadid. *SKANIKA: Sistem Komputer Dan Teknik Informatika*, *7*(2), 204–213. https://doi.org/10.36080/skanika.v7i2.3200