EJECTS: Journal Computer, Technology and Informations System
Vol.4 No.1 September 2024, P-ISSN: 2988-5477 E-ISSN: 2988-4446
LPPM Universitas Darwan Ali

Membangun Model Prediksi Yang Robust: Penanganan Feature Leakage dalam Analisis Dampak Media Sosial Terhadap Prestasi Akademik Siswa

Muhammad Riansyahputra

Program Studi : Sistem Informasi, Universitas Darwan Ali
Email : riankynn@gmail.com

ABSTRACT— The development and evaluation of a robust data mining model was undertaken to forecast the influence of social media activity on student academic outcomes. A critical step in the analytical process involved identifying and resolving a feature leakage issue, which led to the exclusion of two features with deterministic connections (Addicted_Score and Mental_Health_Score) to ensure a realistic and generalizable model. The efficacy of three distinct classifiers—Random Forest, Decision Tree, and k-NN—was subsequently assessed using a 10-fold Cross Validation technique to guarantee reliable findings. After these adjustments, the Random Forest algorithm demonstrated superior performance, achieving 99.86% accuracy and a 99.85% F1-Score. This outcome confirms that significant predictive power is retained within other behavioral variables, such as usage duration and sleep patterns. A trustworthy predictive model was successfully produced, underscoring the critical importance of meticulous data analysis for practical, real-world applications.

Keywords— Data Mining, Academic Achievement, Random Forest, Cross Validation, Data Leakage.

ABSTRAK— Pengembangan dan evaluasi sebuah model *Data Mining* yang andal dilakukan untuk memprediksi dampak penggunaan media sosial terhadap prestasi akademik siswa. Proses analisisnya diawali dengan identifikasi dan penanganan masalah *feature leakage*, di mana fitur-fitur dengan hubungan deterministik (Addicted_Score dan Mental_Health_Score) dihilangkan demi membangun model yang realistis dan dapat digeneralisasi. Tiga algoritma klasifikasi, yaitu Random Forest, Decision Tree, dan k-NN, kemudian dievaluasi kinerjanya menggunakan metode *Cross Validation* 10-folds untuk memastikan keandalan hasil. Setelah dilakukan penyesuaian, model Random Forest menunjukkan kinerja terbaik dengan akurasi 99,86% dan F1-Score 99,85%. Hasil ini membuktikan bahwa pola prediktif yang kuat masih dapat ditemukan dari variabel perilaku yang tersisa, seperti durasi penggunaan dan jam tidur. Sebuah model prediktif yang andal pada akhirnya berhasil dibangun, sekaligus menyoroti pentingnya analisis data yang kritis untuk aplikasi dunia nyata.

Kata Kunci— *Data Mining*, Pencapaian Akademik, Random Forest, *Cross Validation*, Kebocoran Fitur.

I. PENDAHULUAN

Era digital ditandai oleh integrasi media sosial yang mendalam ke dalam rutinitas harian, terutama bagi kalangan pelajar. Platform seperti Instagram, TikTok, dan Twitter telah menjadi ekosistem digital utama bagi generasi muda untuk berinteraksi, mencari informasi, dan membentuk identitas, menjadikan mereka salah satu kelompok pengguna paling intensif secara global.

Seiring manfaatnya, penggunaan yang tinggi ini memunculkan fenomena kecanduan media sosial. Kondisi ini didefinisikan sebagai adiksi perilaku yang ditandai oleh dorongan kompulsif untuk terus menggunakan media sosial, meskipun menyadari adanya konsekuensi negatif [1]. Mekanisme psikologisnya serupa dengan adiksi zat, di mana validasi sosial instan seperti "likes" dan komentar melepaskan dopamin dan menciptakan siklus ketergantungan yang sulit diputus.

Berbagai penelitian telah mendokumentasikan dampak negatif dari kecanduan ini. Studi secara konsisten menunjukkan korelasi kuat antara penggunaan media sosial berlebih dengan memburuknya kesehatan mental, seperti peningkatan gejala kecemasan dan depresi [2]. Lebih lanjut, dampak ini juga merambah ke aspek fisiologis, terutama gangguan pola tidur yang signifikan akibat paparan layar sebelum tidur, yang menyebabkan kelelahan kronis di siang hari [3].

Kondisi kesehatan mental yang terganggu dan kurangnya istirahat berkualitas tersebut secara langsung menciptakan fondasi yang rapuh bagi keberhasilan akademik. Siswa yang mengalami tekanan psikologis dan kelelahan fisik akan kesulitan memusatkan konsentrasi dan mengalami penurunan motivasi belajar. Jalur inilah yang menjadi hipotesis sentral di mana kecanduan media sosial menjadi anteseden bagi penurunan prestasi.

Bidang *Educational Data Mining* (EDM) menawarkan metode untuk menganalisis data siswa dan memprediksi hasil akademik mereka. Tinjauan literatur menunjukkan bahwa data demografis dan perilaku siswa merupakan prediktor yang umum digunakan untuk membangun model prediksi yang akurat dan menyediakan dasar untuk intervensi yang tepat sasaran [4].

Secara spesifik, algoritma yang digunakan dalam penelitian ini telah terbukti efektif dalam konteks EDM. Penelitian sebelumnya telah berhasil menerapkan k-NN untuk klasifikasi siswa [5], menggunakan Decision Tree

untuk mengidentifikasi dan menjelaskan secara transparan dan memanfaatkan kekuatan ansambel dari Random Forest untuk memprediksi siswa yang berisiko (*at-risk*) dengan akurasi tinggi [6].

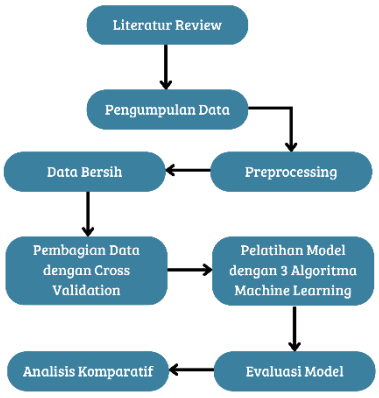
Meskipun hubungan umum antara media sosial dan prestasi akademik telah diketahui, masih ada kebutuhan untuk model prediktif yang kuat dan dapat berfungsi sebagai alat praktis dengan tujuan mengisi kesenjangan tersebut. Kemudian membandingkan tiga algoritma klasifikasi secara sistematis, untuk menjawab pertanyaan: "Apakah kita dapat memprediksi secara andal bahwa prestasi akademik seorang siswa berisiko terpengaruh, hanya dengan menganalisis data demografis dan pola perilaku mereka di media sosial?"

METODOLOGI PENELITIAN

Metodologi penelitian diuraikan secara komprehensif guna menjamin transparansi dan replikabilitas. Penjelasan mencakup semua fase yang dilakukan, mulai dari perancangan desain, teknik pengumpulan data, hingga pendekatan dalam mengevaluasi model.

A. Desain Penelitian

Artikel dibangun menggunakan pendekatan kuantitatif dengan menerapkan teknik-teknik data mining untuk klasifikasi dan prediksi. Alur kerja penelitian dirancang secara sistematis untuk memastikan proses yang logis dan terstruktur. Tahapan-tahapan utama dalam penelitian ini mencakup pengumpulan data, pra-pemrosesan data, pembagian data, pelatihan model menggunakan tiga algoritma yang berbeda, evaluasi kinerja model, dan analisis komparatif untuk penarikan kesimpulan.



Gambar. 1 Bagan Penelitian

B. Sumber dan Deskripsi Data

Data yang digunakan dalam penelitian ini adalah dataset publik berjudul "Students' Social Media Addiction" yang bersumber dari platform Kaggle.[16] Dataset ini merupakan hasil survei yang dirancang untuk mengumpulkan informasi mengenai kebiasaan penggunaan media sosial di kalangan siswa dan potensi dampaknya. Dataset ini terdiri dari 705 baris data (responden) dan 13 kolom (variabel), yang memberikan gambaran komprehensif mengenai demografi, perilaku, dan persepsi siswa terkait media sosial. Variabel target yang akan diprediksi dalam penelitian ini adalah Affects_Academic_Performance, yang merupakan variabel biner ('Yes' atau 'No'). Deskripsi rinci untuk setiap variabel disajikan pada Tabel I.

Tabel I Deskripsi Dataset

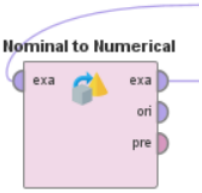
Nama Variabel	Tipe Data	Deskripsi
Student_ID	Integer	Identifier unik responden
Age	Integer	Usia responden dalam tahun
Gender	Categorical	Jenis kelamin (“Male” atau “Female”)
Academic_Level	Categorical	Jenjang akademik (High School, Undergraduate, Graduate)
Country	Categorical	Negara asal responden
Avg_Daily_Usage_Hours	Float	Rata-rata jam penggunaan media sosial per hari
Most_Used_Platform	Categorical	Platform yang paling sering digunakan
Affects_Academic_Performance	Boolean	Dampak yang dirasakan terhadap prestasi akademik (Yes/No). Ini adalah label.
Sleep_Hours_Per_Night	Float	Rata-rata jam tidur per malam
Mental_Health_Score	Integer	Skor kesehatan mental yang dirasakan (1=buruk hingga 10=sangat baik)
Relationship_Status	Categorical	Status hubungan (Single, In Relationship, Complicated)
Conflicts_Over_Social_Media	Integer	Jumlah konflik dalam hubungan yang disebabkan oleh media sosial
Addicted_Score	Integer	Skor kecanduan media sosial (1=rendah hingga 10=tinggi)

C. Pra-Pemrosen Data

Sebelum data dapat digunakan untuk melatih model, serangkaian langkah pra-pemrosesan dilakukan untuk mengolah data mentah menjadi format yang sesuai untuk algoritma *machine learning*.

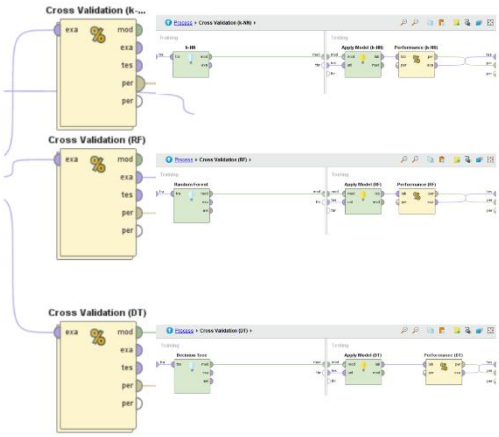
1. Penanganan Data Kategorikal: Variabel-variabel kategorikal seperti Gender, Academic_Level, Country, Most_Used_Platform, dan Relationship_Status tidak dapat diproses secara langsung oleh sebagian besar algoritma. Oleh karena itu, teknik *One-Hot Encoding* diterapkan untuk mengubah setiap kategori dalam variabel tersebut menjadi kolom biner baru. Variabel target, Affects_Academic_Performance, yang bersifat

biner, diubah menjadi format numerik, di mana 'Yes' direpresentasikan sebagai 1 dan 'No' sebagai 0.



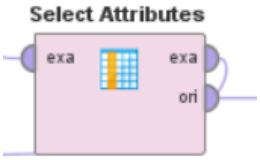
Gambar. 2 Operator Nominal to Numerical

2. Evaluasi Model dengan Validasi Silang: Untuk evaluasi model yang kuat dan dapat diandalkan, penelitian ini mengadopsi metode Cross Validation (k-fold Cross Validation) dengan k=10. Metode ini dipilih sebagai pengganti pembagian data tunggal (*single split data*) untuk mengatasi potensi bias dan varians dalam evaluasi model. Tidak seperti *split data* yang hanya menguji model pada satu set data pengujian yang tetap, Validasi Silang melakukan pengujian sebanyak 10 kali, di mana setiap bagian data secara bergantian berfungsi sebagai data pengujian. Hal ini memastikan bahwa evaluasi kinerja tidak bergantung pada keberuntungan pembagian data secara acak dan memberikan estimasi yang lebih stabil dan representatif terhadap kemampuan generalisasi model pada data baru.



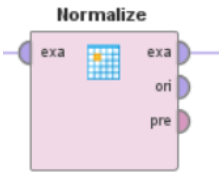
Gambar. 4 Operator Cross Validation

3. Penanganan Kebocoran Data (Data Leakage): Selama analisis eksplorasi awal, teridentifikasi adanya hubungan deterministik yang sempurna antara beberapa fitur (Addicted_Score dan Mental_Health_Score) dengan variabel target. Fenomena ini, yang dikenal sebagai *feature leakage*, dapat menghasilkan model dengan kinerja yang terlalu optimis dan tidak realistis. Untuk membangun model yang lebih robust dan dapat digeneralisasi, fitur-fitur yang menyebabkan kebocoran data ini secara sengaja **dihilangkan** dari dataset sebelum proses pemodelan lebih lanjutPenanganan Kebocoran Data (Data Leakage): Selama analisis eksplorasi awal, teridentifikasi adanya hubungan deterministik yang sempurna antara beberapa fitur (Addicted_Score dan Mental_Health_Score) dengan variabel target. Fenomena ini, yang dikenal sebagai *feature leakage*, dapat menghasilkan model dengan kinerja yang terlalu optimis dan tidak realistis. Untuk membangun model yang lebih robust dan dapat digeneralisasi, fitur-fitur yang menyebabkan kebocoran data ini secara sengaja **dihilangkan** dari dataset sebelum proses pemodelan lebih lanjut



Gambar. 3 Operator Select Attributes

4. Penskalaan Fitur (Feature Scaling): Variabel numerik dalam dataset, seperti Age, Avg_Daily_Usage_Hours & Sleep_Hours_Per_Night, memiliki rentang nilai yang berbeda-beda. Untuk mencegah variabel dengan rentang nilai yang lebih besar secara tidak adil mendominasi proses pelatihan, terutama pada algoritma yang sensitif terhadap jarak seperti k-NN, teknik penskalaan fitur diterapkan. Penelitian ini menggunakan **Standardisasi Z-score**, yang mengubah setiap fitur numerik sehingga memiliki rata-rata (mean) 0 dan standar deviasi 1. Dengan membawa semua fitur ke skala yang sama, metode ini memastikan bahwa setiap variabel memberikan kontribusi yang setara terhadap hasil model, berdasarkan pola informasinya, bukan karena skala arbitrer dari pengukurannya



Gambar. 5 Operator Normalize

D. Model Klasifikasi

Penelitian ini membandingkan tiga algoritma klasifikasi yang berbeda untuk memprediksi dampak media sosial terhadap prestasi akademik.

1. *k-Nearest Neighbors (k-NN)*: k-NN adalah algoritma pembelajaran yang sederhana namun efektif, bersifat non-parametrik dan berbasis instans. Prinsip kerjanya adalah mengklasifikasikan sebuah titik data baru berdasarkan kelas mayoritas dari 'k' tetangga terdekatnya dalam ruang fitur. Jarak antara titik data dihitung menggunakan metrik jarak Euclidean, dan pemilihan nilai 'k' yang optimal merupakan faktor kunci dalam menentukan kinerja model ini.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Rumus tersebut adalah Jarak Euclidean (Euclidean Distance), yang berfungsi untuk mengukur jarak garis lurus antara dua titik data, p dan q. Perhitungannya dilakukan dengan cara menjumlahkan kuadrat dari selisih setiap pasang fitur yang bersesuaian, lalu mengambil akar kuadrat dari totalnya. Dalam machine learning, metode ini merupakan inti dari algoritma K-NN untuk menentukan seberapa "dekat" suatu data dengan tetangga-tetangganya secara akurat

2. *Decision Tree*: adalah salah satu algoritma *supervised learning* yang paling intuitif dan mudah diinterpretasikan. Cara kerjanya meniru proses pengambilan keputusan manusia, yang divisualisasikan dalam struktur seperti pohon. Proses dimulai dari akar (root node), yang merupakan fitur terpenting yang paling baik dalam memisahkan data. Dari akar ini, data dibagi menjadi beberapa cabang (branch) berdasarkan aturan atau kriteria keputusan tertentu.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

Rumus Gain Ratio menyempurnakan Information Gain dengan menyeimbangkan keuntungan informasi dengan tingkat kerumitan sebuah fitur (SplitInfo). Tujuannya adalah untuk mengatasi bias terhadap fitur yang punya banyak kategori, dengan cara "menghukum" fitur yang terlalu kompleks. Hasilnya, pohon keputusan memilih fitur yang paling informatif sekaligus efisien.

3. *Random Forest*: Random Forest adalah metode pembelajaran ansambel yang sangat kuat dan fleksibel, yang bekerja dengan membangun sejumlah besar pohon keputusan (*decision trees*) pada saat pelatihan. Prediksi akhir dari model ini ditentukan oleh kelas yang paling sering muncul (modus) dari semua prediksi pohon keputusan individu [7]. Keunggulan utama Random Forest meliputi akurasi yang tinggi, ketahanan terhadap *overfitting*, dan kemampuannya untuk mengukur tingkat kepentingan relatif dari setiap fitur dalam membuat prediksi [8].

E. Metrik Evaluasi Kinerja

Kinerja dari ketiga model klasifikasi dievaluasi menggunakan serangkaian metrik standar yang berasal dari *confusion matrix*. *Confusion matrix* adalah sebuah tabel yang merangkum hasil prediksi dengan membandingkannya dengan kelas aktual, yang terdiri dari empat komponen: *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN). Metrik-metrik yang digunakan adalah sebagai berikut:

1. Akurasi (Accuracy): Mengukur proporsi prediksi yang benar secara keseluruhan. Dihitung dengan rumus:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Presisi (Precision): Mengukur proporsi prediksi positif yang benar-benar positif. Metrik ini penting ketika biaya dari False Positive tinggi. Dihitung dengan rumus:

$$Presisi = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity): Mengukur proporsi kasus positif aktual yang berhasil diidentifikasi dengan benar oleh model. Metrik ini krusial ketika biaya dari False Negative tinggi. Dihitung dengan rumus:

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score: Merupakan rata-rata harmonik dari Presisi dan Recall, memberikan skor tunggal yang menyeimbangkan kedua metrik tersebut. Metrik ini sangat berguna ketika terdapat ketidakseimbangan kelas dalam data. Dihitung dengan rumus:

$$F1 - Score = 2\,x\,\frac{Presisi\ x\ Recall}{Presisi + Recall}$$

II. DESAIN, HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari eksperimen yang telah dilakukan, diikuti dengan analisis dan pembahasan mendalam terhadap temuan tersebut. Hasil kuantitatif dari kinerja model disajikan terlebih dahulu, kemudian diinterpretasikan dalam konteks masalah penelitian.

A. Hasil Penelitian

Setelah melalui proses pra-pemrosesan data yang cermat, termasuk penghapusan fitur yang menyebabkan kebocoran data, ketiga model klasifikasi dievaluasi menggunakan metode Cross Validation 10 folds. Hasil kinerja rata-rata dari ketiga model tersebut dirangkum secara komparatif dalam Tabel II.

Tabel II Hasil Kinerja Model

Model	Accuracy	Precision	Recall	F1- score
Random forest	99.86%	99.80%	99.89%	99.85%
Decision tree	99.57%	99.41%	99.67%	98.92%
k-NN	98.58%	98.38%	98.55%	98.45%

Berdasarkan Tabel II, terlihat bahwa model **Random Forest** menunjukkan kinerja superior secara keseluruhan, meskipun Regresi Logistik juga memberikan

hasil yang sangat kompetitif. Random Forest mencapai akurasi tertinggi sebesar 99.86% dan F1-Score tertinggi sebesar 99.85%, menjadikannya model yang paling seimbang dan andal dalam memprediksi dampak akademik berdasarkan fitur-fitur yang tersisa.

Untuk memberikan gambaran yang lebih detail mengenai kinerja model terbaik, *confusion matrix* gabungan dari hasil Cross Validation untuk model Random Forest disajikan pada Gambar 6. Matriks ini secara visual menunjukkan distribusi prediksi yang benar dan salah dari keseluruhan data.

accuracy: 99.86% +/- 0.45% (micro average: 99.86%)

	true Yes	true No	class precision
pred. Yes	452	0	100.00%
pred. No	1	252	99.60%
class recall	99.78%	100.00%	

Gambar. 7 Confusion Matrix RF

Dengan dihilangkannya fitur-fitur yang menyebabkan kebocoran data, model Random Forest kini mengandalkan kombinasi fitur lain untuk membuat prediksi. Gambar 7 menampilkan peringkat kepentingan fitur (*feature importance*) yang dihasilkan oleh model Random Forest final dengan Operator Weight by Gini Index. Hasil ini memberikan wawasan tentang variabel mana yang paling berpengaruh dalam memprediksi dampak akademik setelah penyesuaian. Terlihat bahwa Conflicts_Over_Social_Media dan Avg_Daily_Usage_Hours menjadi dua prediktor teratas.

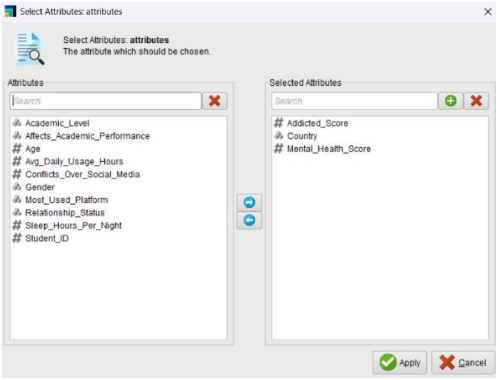
Sleep_Hours_Per_Night	0.140
Avg_Daily_Usage_Hours	0.168
Conflicts_Over_Social_Media	0.451

Gambar. 8 Atribut paling berpengaruh

B. Pembahasan

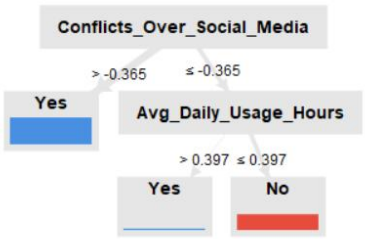
Proses pemodelan dalam penelitian ini melalui sebuah iterasi penting yang secara signifikan meningkatkan validitas dan reliabilitas hasil.

1. *Penanganan Kebocoran Data dan Implikasinya*: Pada analisis awal, ditemukan bahwa model Random Forest mampu mencapai akurasi 100% secara konsisten, bahkan dengan metode Cross Validation. Investigasi lebih lanjut mengungkapkan bahwa fenomena ini disebabkan oleh *feature leakage*, di mana fitur Addicted_Score dan Mental_Health_Score memiliki hubungan deterministik yang sempurna dengan variabel target. Hal ini mengindikasikan bahwa dataset kemungkinan besar dibuat dengan aturan-aturan kaku, bukan mencerminkan variabilitas data dunia nyata. Untuk membangun model yang prediktif dan bukan sekadar mereplikasi aturan, kedua fitur tersebut dihilangkan. Langkah ini krusial untuk memastikan model yang dihasilkan benar-benar "belajar" dari pola yang kompleks, bukan mengambil "jalan pintas".



Gambar. 6 Buang atribut yang tidak di gunakan

2. *Visualisasi Decision Tree*: Decision Tree menggambarkan alur prediksi untuk hasil "Yes" atau "No" yang dimulai dari fitur paling signifikan, yaitu Conflicts_Over_Social_Media. Jika nilai fitur ini di atas -0.365, model dengan keyakinan tinggi langsung menyimpulkan hasilnya adalah "Yes", yang ditandai oleh bar biru solid. Namun, jika nilainya lebih rendah atau sama dengan -0.365, model akan memeriksa kondisi kedua pada fitur Avg_Daily_Usage_Hours. Di jalur ini, jika penggunaan harian rata-rata di atas 0.397, hasilnya tetap "Yes", sedangkan jika di bawah atau sama dengan 0.397, prediksinya berubah menjadi "No" yang direpresentasikan oleh bar merah. Secara keseluruhan, model ini menunjukkan tiga jalur prediksi yang berbeda, dengan Conflicts_Over_Social_Media sebagai faktor penentu utamanya.



Gambar. 9 Decision Tree

3. *Interpretasi Kinerja Model Final*: Setelah penyesuaian, Random Forest tetap menjadi model dengan kinerja terbaik (Akurasi 99.86%, F1-Score 99.85%), diikuti sangat dekat oleh Decision Tree (Akurasi 99.57%, F1-Score 98.92%). Kinerja yang sangat tinggi ini, bahkan setelah menghilangkan prediktor terkuat, menunjukkan bahwa masih terdapat sinyal prediktif yang sangat kuat dari kombinasi fitur-fitur yang tersisa. Keunggulan Random Forest dapat diatribusikan pada kemampuannya menangani interaksi non-linear antar variabel, sementara Decision Tree memberikan hasil yang sangat baik dengan model yang lebih sederhana dan mudah diinterpretasikan.

4. *Analisis Faktor Prediktif Utama (Final)*: Analisis kepentingan fitur dari model final menyoroti Avg_Daily_Usage_Hours dan Conflicts_Over_Social_Media sebagai faktor paling krusial. Ini adalah temuan yang sangat logis dan sejalan dengan literatur. Durasi penggunaan media sosial dan tingkat konflik yang disebabkan dari media sosial yang tinggi secara langsung memengaruhi kemampuan kognitif dan fokus siswa. Ini menunjukkan bahwa perilaku dasar sehari-hari memiliki daya prediksi yang sangat kuat terhadap persepsi kinerja akademik.
5. *Implikasi Praktis dan Keterbatasan*: Model final yang dikembangkan memiliki potensi besar sebagai sistem peringatan dini yang realistis. Dengan fokus pada variabel perilaku yang dapat diamati seperti durasi penggunaan media sosial, jumlah konflik yang disebabkan oleh media sosial dan pola tidur, institusi pendidikan dapat merancang intervensi yang lebih praktis dan tidak terlalu intrusif. Keterbatasan utama tetap pada sifat data yang dilaporkan sendiri (*self-reported*). Meskipun model ini sangat baik dalam memprediksi *persepsi* mahasiswa, validasi lebih lanjut terhadap data kinerja akademik objektif (misalnya, IPK) akan menjadi langkah berikutnya yang ideal untuk memperkuat temuan ini.

III. KESIMPULAN

Penelitian ini berhasil mengembangkan dan memvalidasi model *data mining* untuk memprediksi dampak penggunaan media sosial terhadap prestasi akademik siswa, dengan melalui proses analisis kritis untuk memastikan hasil yang robust dan realistis.

Kesimpulan utama dari penelitian ini adalah, bahkan setelah menghilangkan fitur-fitur yang menyebabkan kebocoran data, dimungkinkan untuk memprediksi persepsi siswa mengenai dampak akademik dengan akurasi yang sangat tinggi. Model Random Forest terbukti menjadi yang paling efektif, dengan mencapai akurasi 99.86% dan F1-Score 99.85% melalui metode Cross Validation 10-folds. Kinerja ini menunjukkan bahwa terdapat pola yang sangat kuat dan dapat dipelajari dari data perilaku siswa yang tersisa.

Faktor-faktor prediktif yang paling signifikan dalam model final adalah rata-rata durasi penggunaan harian (Avg_Daily_Usage_Hours) dan jumlah konflik dalam hubungan yang disebabkan oleh media sosial (Conflicts_Over_Social_Media). Temuan ini menggarisbawahi bahwa variabel perilaku dasar memiliki daya prediksi yang sangat kuat, lebih dari sekadar data demografis. Ini mengimplikasikan bahwa intervensi yang berfokus pada manajemen waktu digital dan promosi kebiasaan tidur yang sehat dapat menjadi strategi yang sangat efektif bagi institusi pendidikan.

Proses penelitian ini juga menyoroti pentingnya analisis data yang cermat untuk mengidentifikasi dan menangani anomali seperti *feature leakage*. Dengan mengatasi masalah ini, model yang dihasilkan tidak hanya akurat secara statistik tetapi juga lebih valid secara konseptual dan dapat diandalkan untuk aplikasi di dunia nyata sebagai sistem peringatan dini bagi konselor, dosen, dan administrator pendidikan.

IV. REFERENSI

[1] S. Abdulsalim *et al.*, “Evaluation of Social Media Addiction and Its Relationship with Anxiety and Academic Performance Among Medical and Non-Medical Students: A Cross-Sectional Study from Saudi Arabia,” *Healthc.*, vol. 13, no. 3, 2025, doi: 10.3390/healthcare13030295.

[2] A. Mufidah, U. Rohman, and S. Ismail, “Pengaruh Intensitas Penggunaan Media Sosial Terhadap Kesehatan Mental Mahasiswa UIN Sunan Gunung Djati Bandung,” *J. Psikol. Insight*, vol. 9, no. 1, pp. 45–56, 2025.

[3] I. H. Utami, N. A. Shifa, and N. Rukiah, “Durasi Penggunaan Media Sosial dengan Kualitas Tidur dan Kestabilan Emosi pada Mahasiswa Keperawatan Tahun 2023,” *Vitalitas Medis J. Kesehat. dan Kedokt.*, vol. 1, no. 2, pp. 81–94, 2024, [Online]. Available: https://journal.lpkd.or.id/index.php/ViMed/article/view/140

[4] S. Alturki, N. Alturki, and H. Stuckenschmidt, “Using Educational Data Mining To Predict Students’ Academic Performance For Applying Early Interventions,” *J. Inf. Technol. Educ. Innov. Pract.*, vol. 20, pp. 121–137, 2021, doi: 10.28945/4835.

[5] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, “Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa,” *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 2, pp. 118–127, 2021, doi: 10.31294/ijcit.v6i2.10438.

[6] R. A. Mrg and M. S. Hasibuan, “Best Student Classification using Ensemble Random Forest Method,” *Sistemasi*, vol. 13, no. 3, p. 1188, 2024, doi: 10.32520/stmsi.v13i3.4101.

[7] A. Ramadhan, B. Susetyo, and Indahwati, “Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan,” *J. Pendidik. dan Kebud.*, vol. 4, no. 2, pp. 169–182, 2019, doi: 10.24832/jpnk.v4i2.1327.

[8] M. Adnan *et al.*, “Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models,” *IEEE Access*, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[9] C. Chaka, “Educational data mining, student academic performance prediction, prediction

methods, algorithms and tools: an overview of reviews," *J. E-Learning Knowl. Soc.*, vol. 18, no. 2, pp. 58–69, 2022, doi: 10.20368/1971-8829/1135578.

[10] S. Sperandei, "Understanding logistic regression analysis," *Biochem. Medica*, vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

[11] R. Dong, D. Yuan, X. Wei, J. Cai, Z. Ai, and S. Zhou, "Exploring the relationship between social media dependence and internet addiction among college students from a bibliometric perspective," *Front. Psychol.*, vol. 16, no. March, pp. 1–20, 2025, doi: 10.3389/fpsyg.2025.1463671.

[12] T. Margareth, "Hubungan penggunaan media sosial dengan kualitas tidur pada remaja di smk negeri 2 binjai tahun 2023," *J. Keperawatan Sisthana*, vol. 8, no. 2, pp. 47–60, 2023, doi: 10.55606/sisthana.v8i2.562.

[13] D. Olivia, "Penerapan Algoritma K-Nearest Neighbor (KNN) Untuk Ketepatan Waktu Lulus Mahasiswa," pp. i–57, 2024.

[14] J. O. Esieboma, "University Students ' Mental Health in the Age of Social Media : A Sociological Perspective," vol. 8, no. 5, pp. 73–82, 2024, doi: 10.56201/ijmepr.v8.no5.2024.pg73.82.

[15] Y. Tampil, H. Komaliq, and Y. Langi, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado," *d'CARTESIAN*, vol. 6, no. 2, p. 56, 2017, doi: 10.35799/dc.6.2.2017.17023.

[16] E. Yildirim Demirdöğen *et al.*, "Social media addiction, escapism and coping strategies are associated with the problematic internet use of adolescents in Türkiye: a multi-center study," *Front. Psychiatry*, vol. 15, no. February, pp. 1–10, 2024, doi: 10.3389/fpsyt.2024.1355759.

[17] N. Nurmalitasari and E. Purwanto, "Prediksi Performa Mahasiswa Menggunakan Model Regresi Logistik," *J. Deriv. J. Mat. dan Pendidik. Mat.*, vol. 9, no. 2, pp. 145–152, 2022, doi: 10.31316/jderivat.v9i2.2639.