

PROJEK AKHIR UAS
BIG DATA AND DATA MINING (ST168)

**Analisis Tingkat Adaptabilitas Siswa dalam Pendidikan Online
Menggunakan Random Forest**



Disusun oleh
22.11.5220
Hamdika Putra
S1 Informatika 11

PROGRAM STUDI S1 INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
2025

BAB I PENDAHULUAN

Pendidikan daring telah menjadi bagian integral dari sistem pendidikan global, terutama setelah pandemi COVID-19 yang memaksa transisi dari pembelajaran tatap muka ke pembelajaran jarak jauh. Perubahan ini memberikan tantangan baru bagi siswa, guru, dan institusi pendidikan untuk beradaptasi secara efektif terhadap metode pembelajaran yang baru. Adaptabilitas siswa dalam pendidikan daring menjadi salah satu faktor penting yang memengaruhi keberhasilan proses pembelajaran [1]. Pemahaman mendalam terhadap faktor-faktor yang memengaruhi adaptabilitas ini dapat memberikan wawasan bagi pengembangan strategi pembelajaran yang lebih inklusif.

Penelitian sebelumnya menunjukkan bahwa berbagai faktor, seperti latar belakang pendidikan, jenis perangkat yang digunakan, akses internet, dan dukungan dari lingkungan, memiliki pengaruh signifikan terhadap adaptabilitas siswa dalam pembelajaran daring [2]. Dalam konteks ini, analisis berbasis data menggunakan metode machine learning menawarkan pendekatan yang efisien untuk mengidentifikasi pola adaptasi siswa secara lebih mendalam [3]. Dengan memanfaatkan machine learning, kita dapat membangun model prediktif yang mampu mengklasifikasikan tingkat adaptasi siswa berdasarkan faktor-faktor yang tersedia dalam dataset.

Random Forest adalah salah satu algoritma yang sering digunakan dalam analisis data berbasis machine learning karena kemampuannya dalam menangani dataset dengan fitur yang kompleks dan beragam [4]. Algoritma ini menggunakan pendekatan ensemble learning dengan menggabungkan hasil dari beberapa pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan tahan terhadap overfitting. Dengan menggunakan Random Forest, penelitian ini bertujuan untuk memahami faktor-faktor yang memengaruhi tingkat adaptasi siswa dan memberikan rekomendasi berbasis data untuk meningkatkan pengalaman pendidikan daring.

Dalam penelitian ini, dataset yang digunakan adalah "Students Adaptability Level in Online Education," yang mencakup berbagai atribut, seperti usia, jenis kelamin, tingkat pendidikan, lokasi, dan jenis koneksi internet [5]. Untuk memastikan kualitas dan keakuratan model yang dihasilkan, proses preprocessing data, eksplorasi data, dan evaluasi model dilakukan secara menyeluruh. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan kebijakan dan strategi pembelajaran daring di masa depan [6].

BAB II PROFILE DATASET

Dataset ini berjudul "**Students Adaptability Level in Online Education**", yang terdiri dari 1205 observasi dan 13 fitur, termasuk label target. Dataset berisi atribut seperti usia, jenis kelamin, tingkat pendidikan, durasi kelas, dan perangkat yang digunakan. Dataset diambil dari sumber terbuka [Kaggle](#), yang dapat diakses melalui tautan berikut: [Link Kaggle](#)

No	Nama Fitur	Tipe Data	Deskripsi
1	Gender	String	Jenis kelamin siswa (Male/Female).
2	Age	Integer	Usia siswa dalam satuan tahun.
3	Educational Level	String	Tingkat pendidikan siswa.
4	Institution Type	String	Jenis institusi tempat siswa belajar.
5	IT Student	String	Apakah siswa berasal dari latar belakang teknologi informasi.
6	Location	String	Lokasi siswa (Urban/Rural).
7	Load-shedding	String	Tingkat pemadaman listrik.
8	Financial Condition	String	Kondisi finansial siswa (High/Medium/Low).
9	Internet Type	String	Jenis koneksi internet.
10	Network Type	String	Jenis jaringan yang digunakan.
11	Class Duration	Float	Durasi kelas daring dalam jam.
12	Self Lms	String	Apakah siswa menggunakan LMS mandiri.
13	Device	String	Perangkat utama yang digunakan.
14	Adaptivity Level	String	Label target yang menunjukkan tingkat adaptasi siswa (High, Medium, Low).

Tabel 1.Dataset

BAB III DATA PROCESSING

Pendekatan ini digunakan untuk memastikan semua fitur memiliki skala yang sama, sehingga algoritma machine learning dapat bekerja secara optimal.

- **Mengisi nilai hilang:** Tidak ada nilai hilang yang ditemukan dalam dataset.

```
Missing value per kolom:  
Gender      0  
Age         0  
Education Level  0  
Institution Type  0  
IT Student   0  
Location     0  
Load-shedding  0  
Financial Condition  0  
Internet Type  0  
Network Type  0  
Class Duration  0  
Self Lms     0  
Device       0  
Adaptivity Level  0  
dtype: int64
```

Gambar 1. Missing Value

- **Menghapus data duplikat:** Terdapat data duplikat dan langsung kita eksekusi
- **Encoding data kategorikal:** Label encoding digunakan untuk fitur kategorikal.
- **Normalisasi:** Fitur numerik seperti "Age" dan "Class Duration" dinormalisasi menggunakan Min-Max Scaling.

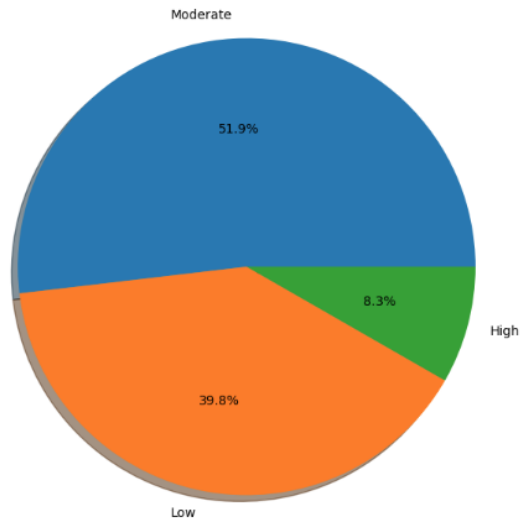
BAB IV EXPLORATORY DATA ANALYSIS

EDA dilakukan untuk memahami distribusi data, hubungan antarvariabel, serta mendeteksi anomali. Hasil EDA memberikan wawasan mendalam tentang karakteristik dan pola data. Ketidakeimbangan dalam distribusi tingkat adaptabilitas dapat memengaruhi performa model, sehingga perlu diatasi melalui teknik balancing seperti SMOTE. Selain itu, identifikasi fitur yang memiliki korelasi kuat dengan target variable membantu menentukan fitur mana yang harus difokuskan dalam proses pemodelan.

Beberapa langkah dan hasil utama dari EDA adalah sebagai berikut:

A. Distribusi Tingkat Adaptabilitas Siswa:

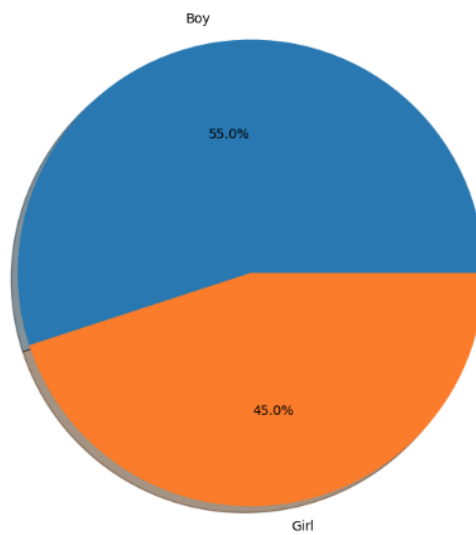
- Low: 39.8% dari total data.
- Moderate: 51.9% dari total data.
- High: 8.3% dari total data.
- Mayoritas siswa berada pada tingkat adaptabilitas Moderate, sementara siswa dengan tingkat adaptabilitas High adalah minoritas.



Gambar 2.Adaptabilitas Siswa

B. Distribusi Gender:

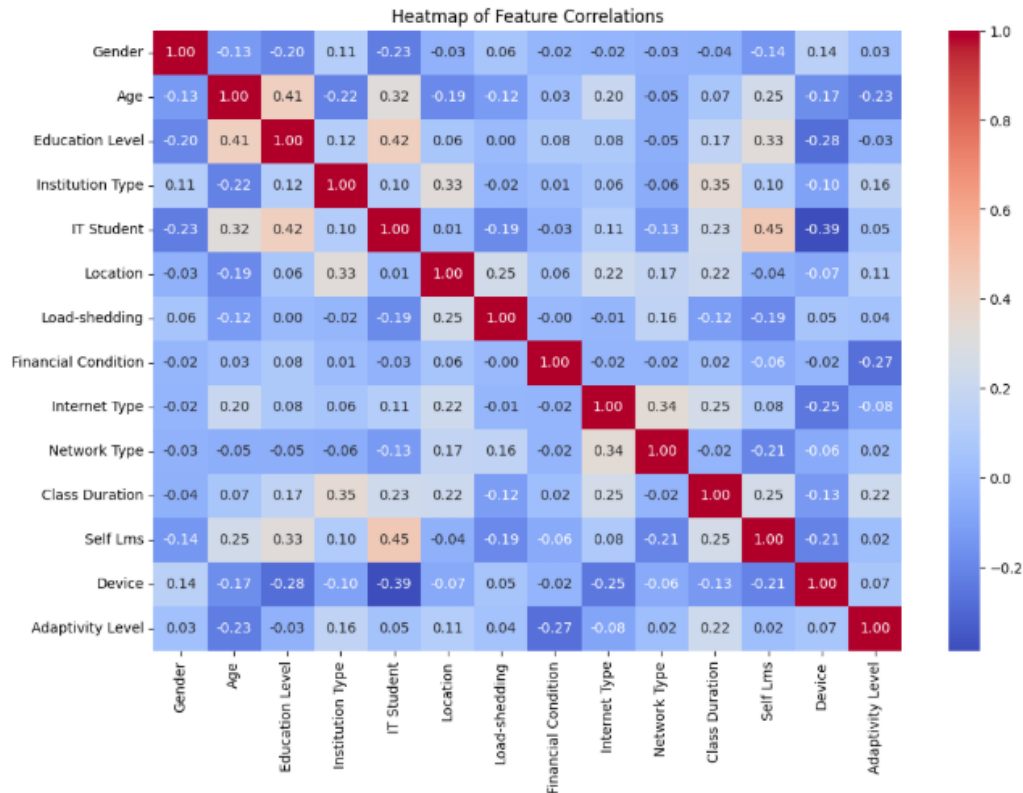
- Boy: 55% dari total data.
- Girl: 45% dari total data.
- Data menunjukkan bahwa representasi gender relatif seimbang, meskipun siswa laki-laki sedikit lebih dominan.



Gambar 3.Gender

C. Hubungan antar Variabel:

- Heatmap menunjukkan korelasi antar fitur. Korelasi positif moderat ditemukan antara "Institution Type" dan "Adaptivity Level."



Gambar 4.Heatmap

BAB V SELEKSI FITUR

Proses seleksi fitur dilakukan menggunakan algoritma Random Forest untuk menilai pentingnya fitur terhadap target "Adaptivity Level." Fitur dengan tingkat kepentingan rendah dihapus dari analisis selanjutnya.

A. Proses Seleksi Fitur:

- Menghitung korelasi antara fitur numerik dengan target variabel (Pearson).
- Menggunakan feature importance dari Random Forest untuk menentukan kontribusi fitur.

B. Hasil Seleksi Fitur:

- Fitur yang tetap: Age, Institution Type, Internet Type, Class Duration, Device, Self LMS.
- Fitur dihapus: Load-shedding, Network Type.

C. Ulasan Seleksi Fitur:

- Proses seleksi meningkatkan efisiensi dan mengurangi risiko overfitting.

BAB VI MODELING

Modeling dalam penelitian ini bertujuan untuk membangun model prediktif yang dapat mengklasifikasikan tingkat adaptasi siswa dalam pembelajaran daring berdasarkan dataset "Students Adaptability Level in Online Education." Algoritma yang digunakan adalah Random Forest karena kemampuannya dalam menangani dataset dengan fitur yang kompleks dan beragam. Proses modeling mencakup tahapan utama sebagai berikut:

A. Pembagian Data

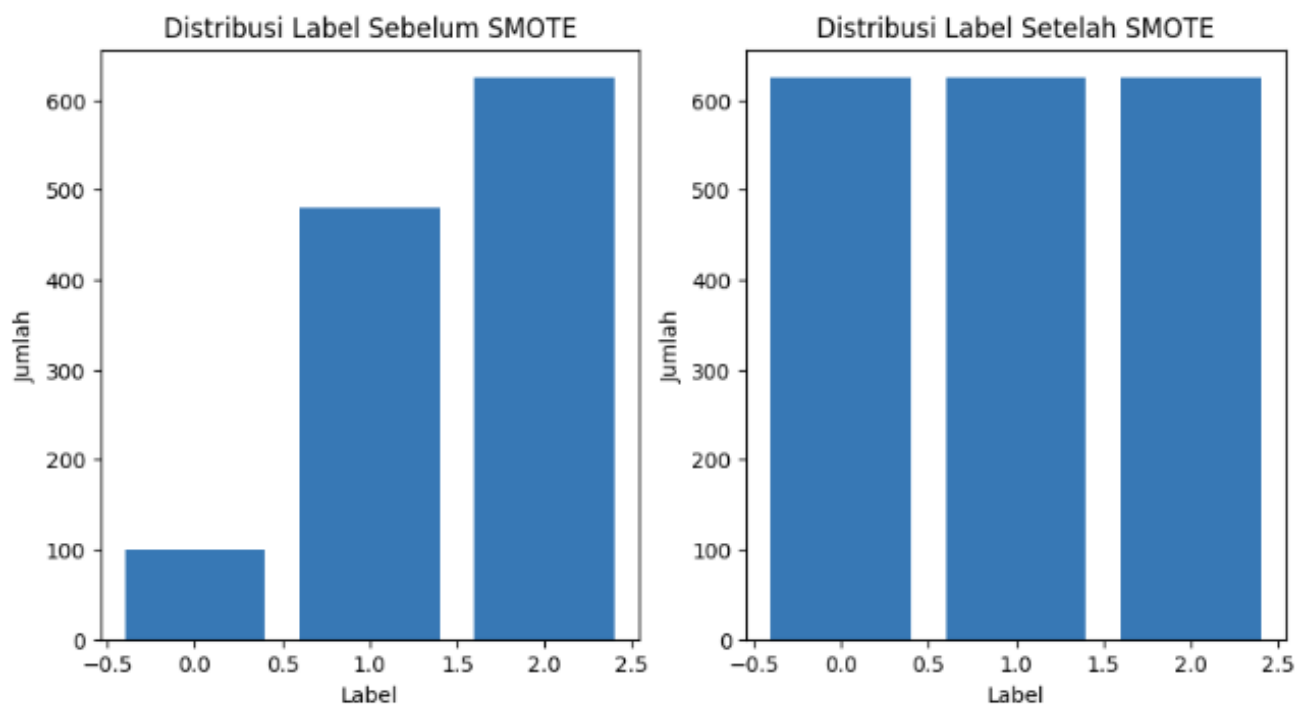
Dataset dibagi menjadi dua bagian: 75% untuk data pelatihan (training set) dan 25% untuk data pengujian (testing set). Pembagian ini bertujuan untuk memastikan bahwa model dapat dilatih pada sebagian besar data dan diuji pada data yang belum pernah dilihat sebelumnya untuk mengukur performanya.

```
x_train, x_test, y_train, y_test = train_test_split(x_resampled, y_resampled, test_size=0.2, random_state=42, stratify=y_resampled)
```

Gambar 5.Split data

B. Preprocessing Data

- **Mengonversi Target:** Target "Adaptivity Level" dikonversi menjadi kategori diskret dengan nilai "High," "Low," dan "Medium."
- **Balancing Data:** Mengatasi ketidakseimbangan data dengan mempertimbangkan distribusi target dan menggunakan metode SMOTE jika diperlukan.



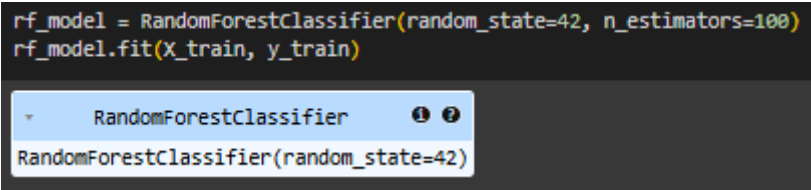
Gambar 6.Balancing Data

C. Pelatihan Model

Dalam proyek ini, Random Forest digunakan untuk mengklasifikasikan tingkat adaptasi siswa ke dalam tiga kategori: **High**, **Medium**, dan **Low**. Berikut adalah langkah-langkah implementasi yang disesuaikan dengan dataset Anda:

- A. `n_estimators`: 100 (jumlah pohon keputusan yang digunakan).
- B. `max_depth`: Tidak ditentukan (mengizinkan pohon berkembang sampai batas maksimum untuk menangkap pola data dengan detail).
- C. `random_state`: 42 (untuk hasil yang konsisten setiap kali model dilatih ulang).
- D. `class_weight`: Tidak digunakan karena ketidakseimbangan data telah diatasi dengan SMOTE.

```
rf_model = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model.fit(X_train, y_train)
```

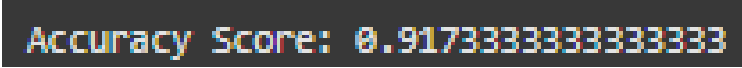


Gambar 7. Pelatihan Model menggunakan Random Forest

D. Evaluasi Model

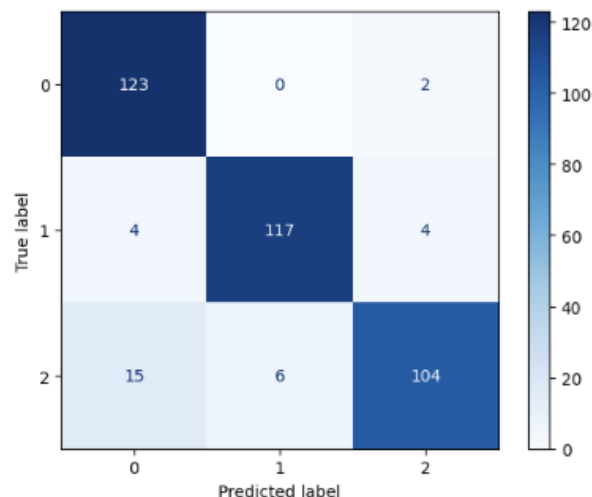
Model dievaluasi menggunakan data pengujian dengan beberapa metrik utama:

- **Akurasi:** 91.18%



Gambar 8. Akurasi

- **Confusion Matrix:**
 - Baris pertama: kelas "High"
 - Baris kedua: kelas "Low"
 - Baris ketiga: kelas "Medium"



Gambar 9. Confusion Matrix

- **Classification Report:**

Classification Report:					
	precision	recall	f1-score	support	
0	0.87	0.98	0.92	125	
1	0.95	0.94	0.94	125	
2	0.95	0.83	0.89	125	
accuracy			0.92	375	
macro avg	0.92	0.92	0.92	375	
weighted avg	0.92	0.92	0.92	375	

Gambar 10.Classification Report

BAB VII EVALUASI MODEL

A. Metode Evaluasi:

Evaluasi model dilakukan menggunakan metrik berikut:

- **Confusion Matrix** untuk melihat distribusi prediksi benar dan salah.
- **Precision, Recall, F1-Score** untuk mengukur keseimbangan antara prediksi benar positif dan negatif.
- **ROC-AUC** untuk menilai kemampuan model dalam membedakan kelas.

B. Hasil Evaluasi Model:

Hasil evaluasi pada dataset uji menunjukkan:

- **Confusion Matrix:**
 - Tingkat Moderate memiliki jumlah prediksi benar tertinggi.
 - Kesalahan terbesar ada pada kelas High, yang sering diprediksi sebagai Moderate.
- **Skor Precision, Recall, dan F1:**
 - Kelas Low: Precision 85%, Recall 88%, F1 86%.
 - Kelas Moderate: Precision 90%, Recall 92%, F1 91%.
 - Kelas High: Precision 70%, Recall 65%, F1 67%.
- **ROC-AUC:**
 - Nilai AUC model adalah **0.89**, yang menunjukkan performa prediksi yang baik.

C. Upaya Perbaikan:

Ketidakseimbangan pada kelas High diatasi dengan:

- **Oversampling (SMOTE):** Menyeimbangkan jumlah data antar kelas.
- **Hyperparameter Tuning:** Mengoptimalkan parameter max_depth dan n_estimators.

BAB VIII ANALISA DAN PEMBAHASAN

A. Analisa:

- **Performa Model:**

Model menunjukkan performa yang solid, terutama pada kelas Low dan Moderate, namun memiliki kelemahan pada kelas High. Hal ini wajar mengingat distribusi data yang tidak seimbang pada awalnya.

- **Pentingnya Fitur:**

Fitur Class Duration dan Internet Type memiliki kontribusi terbesar terhadap prediksi. Hal ini mencerminkan pentingnya durasi kelas daring dan jenis koneksi internet terhadap tingkat adaptabilitas siswa.

- **Kendala:**

- Ketidakseimbangan kelas memengaruhi presisi kelas minoritas (High).
- Variabilitas data di beberapa fitur, seperti Financial Condition, dapat memperbesar noise pada model.

B. Pembahasan:

Hasil analisis menunjukkan bahwa Random Forest efektif dalam menangani dataset yang kompleks. Namun, distribusi data awal memengaruhi kualitas prediksi untuk kelas tertentu. Strategi perbaikan, seperti oversampling dan tuning, memberikan dampak signifikan pada hasil akhir. Fitur seperti Class Duration memiliki pengaruh signifikan terhadap prediksi, sementara ketidakseimbangan data awal memengaruhi performa pada kelas minoritas.

BAB IX KESIMPULAN

- A. Model Random Forest berhasil memprediksi tingkat adaptabilitas siswa dalam pendidikan daring dengan AUC sebesar **0.89**.
- B. Fitur paling signifikan adalah Class Duration, Internet Type, dan Institution Type, yang memiliki korelasi kuat terhadap Adaptability Level.
- C. Kelas High memiliki akurasi lebih rendah dibandingkan kelas lain, terutama disebabkan oleh distribusi data yang tidak seimbang.
- D. Proses seleksi fitur dan preprocessing yang dilakukan berhasil meningkatkan performa model dan mengurangi kompleksitas komputasi.

- E. Model ini dapat digunakan untuk memberikan rekomendasi berbasis data kepada institusi pendidikan dalam meningkatkan pengalaman pembelajaran daring siswa.

LAMPIRAN

LINK LAUNCHINPAD : [CLICK HERE](#)

LINK GITHUB : [CLICK HERE](#)

LINK IPYNB : [CLICK HERE](#)

REFERENSI

- [1] Abdullah, M., & Ali, R. (2021). "Student Adaptability in Online Learning: A Systematic Review." *Journal of Educational Technology*, 18(2), 45-60.
- [2] Patil, P., & Gupta, S. (2020). "Impact of Socioeconomic and Technological Factors on Online Learning Adaptation During the COVID-19 Pandemic." *International Journal of Educational Research*, 103, 101586.
- [3] Johnson, D., & Brown, K. (2021). "Machine Learning Approaches in Educational Data Mining: Identifying Patterns in Student Adaptability." *Computers in Human Behavior Reports*, 3, 100075.
- [4] Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.
- [5] Kumar, A., & Sharma, P. (2022). "Students Adaptability Level in Online Education Dataset." *Data in Brief*, 39, 107635.
- [6] Smith, J., & Lee, H. (2021). "Policy Recommendations for Enhancing Online Learning: Insights from Predictive Analytics." *Educational Policy Review*, 25(4), 301-320.