Dataset metadata

Variables

The dataset includes 27 original features, 3 engineered features (added during preprocessing), and 1 target variable. Below is a detailed description of each variable, including its type, description, and relevance to the Kenyan context.

Original Features (27)

1. **student_id** (Numeric, Integer)
    - **Description**: Unique identifier for each student (1–5000).
    - **Relevance**: Facilitates tracking individual records without influencing analysis.
2. **age** (Numeric, Integer)
    - **Description**: Age of the student (17–30 years).
    - **Relevance**: Captures demographic diversity, influencing study habits and maturity.
3. **gender** (Categorical, Factor)
    - **Description**: Gender of the student ("Male," "Female," "Other").
    - **Relevance**: Reflects gender diversity, potentially affecting academic outcomes.
4. **residency** (Categorical, Factor)
    - **Description**: Urban or rural residency ("Urban," "Rural").
    - **Relevance**: Captures socioeconomic and access differences in Kenyan students.
5. **socioeconomic_status** (Categorical, Factor)
    - **Description**: Socioeconomic status ("Low," "Middle," "High").
    - **Relevance**: Influences access to resources like internet and study materials.
6. **parental_education** (Categorical, Factor)

- **Description**: Highest parental education level ("None," "Primary," "Secondary," "Tertiary").
- **Relevance**: Reflects family background, impacting academic support.

7. **family_income** (Numeric, Float)
    - **Description**: Monthly family income in Kenyan Shillings (KES), mean ~25,000 KES.
    - **Relevance**: Key indicator of financial resources, affecting student performance.
8. **distance_to_university** (Numeric, Float)
    - **Description**: Distance to university in kilometers (0–100).
    - **Relevance**: Impacts commuting time and attendance, especially for rural students.
9. **study_hours_weekly** (Numeric, Float)
    - **Description**: Hours spent studying per week, mean ~15.
    - **Relevance**: Direct measure of academic effort, a key predictor of performance.
10. **attendance_rate** (Numeric, Float)
    - **Description**: Proportion of classes attended (0.5–1).
    - **Relevance**: Indicates engagement, critical for academic success.
11. **library_usage** (Numeric, Float)
    - **Description**: Hours spent in the library per week, mean ~5.
    - **Relevance**: Reflects resource utilization, varying by campus access.
12. **extracurricular_activities** (Categorical, Factor)
    - **Description**: Participation in activities ("None," "Sports," "Clubs," "Both").
    - **Relevance**: Influences time management and student engagement.
13. **internet_access** (Categorical, Factor)
    - **Description**: Access to internet ("Yes," "No").

- **Relevance**: Critical for online learning, especially in Kenya's digital divide.

14. **device_ownership** (Categorical, Factor)
    - **Description**: Devices owned ("Smartphone," "Laptop," "Both," "None").
    - **Relevance**: Reflects technological access, impacting study capabilities.
15. **previous_grade** (Numeric, Float)
    - **Description**: Previous academic year grade (40–100%).
    - **Relevance**: Indicates prior academic performance, a predictor of current outcomes.
16. **math_score** (Numeric, Float)
    - **Description**: Math exam score, mean ~60.
    - **Relevance**: Key academic metric, influencing overall performance.
17. **science_score** (Numeric, Float)
    - **Description**: Science exam score, mean ~60.
    - **Relevance**: Complements math score in assessing academic ability.
18. **english_score** (Numeric, Float)
    - **Description**: English exam score, mean ~60.
    - **Relevance**: Critical for communication and academic success in Kenya.
19. **study_group_participation** (Categorical, Factor)
    - **Description**: Participation in study groups ("Yes," "No").
    - **Relevance**: Reflects collaborative learning, common in Kenyan universities.
20. **scholarship_status** (Categorical, Factor)
    - **Description**: Scholarship type ("Full," "Partial," "None").
    - **Relevance**: Indicates financial support, affecting student resources.
21. **campus_housing** (Categorical, Factor)
    - **Description**: Housing type ("On-Campus," "Off-Campus").
    - **Relevance**: Impacts commute time and campus engagement.

22. **part_time_job** (Categorical, Factor)
    - **Description**: Part-time job status ("Yes," "No").
    - **Relevance**: Affects time available for study, common among Kenyan students.
23. **commute_time** (Numeric, Float)
    - **Description**: Daily commute time in minutes, mean ~30.
    - **Relevance**: Influences attendance and study time, especially for off-campus students.
24. **sleep_hours** (Numeric, Float)
    - **Description**: Average nightly sleep hours, mean ~7.
    - **Relevance**: Impacts health and academic performance.
25. **stress_level** (Categorical, Factor)
    - **Description**: Self-reported stress level ("Low," "Moderate," "High").
    - **Relevance**: Affects mental health and academic outcomes.
26. **course_load** (Numeric, Float)
    - **Description**: Credit hours taken, mean ~15.
    - **Relevance**: Indicates academic workload, influencing performance.
27. **faculty** (Categorical, Factor)
    - **Description**: Academic faculty ("Engineering," "Business," "Arts," "Sciences," "Education").
    - **Relevance**: Reflects program diversity, potentially affecting performance.

Engineered Features (3)

28. **study_hours_binned** (Categorical, Factor)
    - **Description**: Discretized study_hours_weekly into "Low," "Moderate," "High," "Very High" based on bins (0–10, 10–20, 20–30, 30+ hours).
    - **Relevance**: Simplifies study hours for analysis and modeling, aiding interpretation of study habits.

29. **family_income_binned** (Categorical, Factor)
    - **Description**: Discretized family_income into "Low," "Medium-Low," "Medium-High," "High" based on quartiles of the income distribution.
    - **Relevance**: Enhances interpretability of socioeconomic status, facilitating analysis of its impact on academic performance.
30. **attendance_rate_binned** (Categorical, Factor)
    - **Description**: Discretized attendance_rate into "Low," "Medium," "High" based on bins (0–0.7, 0.7–0.85, 0.85–1).
    - **Relevance**: Simplifies attendance data for modeling, highlighting engagement levels.

Target Variable

- **academic_performance** (Categorical, Factor)
    - **Description**: Multiclass academic performance ("Poor," "Average," "Good," "Excellent"), derived from weighted features (e.g., study hours, scores) with noise.
    - **Relevance**: Primary outcome for analyzing factors influencing student success in Kenyan universities.