■■ Task 4 Report – Feature Engineering

1. Purpose

The main objective of feature engineering was to transform the raw student dataset into a form that is both easier to interpret and suitable for predictive modeling. By carefully creating and refining features, the dataset better reflects important aspects of student behavior and performance, while also being compatible with machine learning requirements.

2. Key Feature Engineering Steps

1. Attendance Categories

Converted attendance_rate (numeric %) into categories: Low (<50%), Medium (50–75%), High (>75%). This simplifies identification of students at risk due to low attendance.

2. GPA Bands

Grouped prior_gpa_10pt (0–10 scale) into bands: Low (<5), Medium (5–8), High (>8). Helps in distinguishing struggling, average, and high-achieving students.

3. Study Hours Binning

Transformed study_hours_per_week into: equal-width bins (0–5, 6–10, 11–15, 16–20, 20+) and quartiles (Q1–Q4). Enables comparisons across light, moderate, and heavy study groups with balanced sizes.

4. Log Transformation

Applied log transformation to study_hours_per_week to reduce skewness and bring distribution closer to normal for modeling.

5. Encoding Categorical Variables

Converted categorical features (gender, city, course_stream, has_internet, device_type, parental_education, scholarship, etc.) into one-hot encoded columns. Ensures compatibility with ML models without introducing false order.

6. Scaling Continuous Features

Rescaled features (age, study_hours_per_week, attendance_rate, prior_gpa_10pt, test_score, fee_paid_inr, etc.) using Standard Scaler (mean=0, std=1) and MinMax Scaler (0–1). Prevents large-value features from dominating and improves training stability.

3. Final Dataset Summary

Rows	200
Columns	248 (after encoding and transformations)
Target Variable	test_score

4. Conclusion

Through careful feature engineering, the raw student dataset was transformed into a structured form that highlights the most important drivers of performance. The dataset is now optimized for machine learning models and remains interpretable for educators and decision-makers alike.