Predicting Student Exam Scores - Notes

Dataset Overview

Columns:

- student_id
- age
- gender
- course
- study_hours
- class_attendance
- internet_access
- sleep_hours
- sleep_quality
- study_method
- facility_rating
- exam_difficulty
- exam_score

Target: exam_score

Dataset Size: 20,000 entries

Data types:

- Float: study_hours, class_attendance, sleep_hours, exam_score
- Integer: student_id, age
- Object (categorical): gender, course, internet_access, sleep_quality, study_method, facility_rating, exam_difficulty

---

General Observations

- **Irrelevant features:** student_id, age, course
- **Highly relevant features:**

    - study_hours: More study hours → higher marks
    - class_attendance: Higher attendance → higher marks, but some variance exists
    - study_method: Efficient methods → higher marks
    - facility_rating: Better facilities → higher marks
- **Moderately relevant features:**
    - sleep_hours: Both too little and too much sleep can reduce performance
    - sleep_quality: Good quality may slightly improve marks
    - exam_difficulty: Harder exams → lower marks
- **Interactions & patterns:**
    - High study_hours + low sleep_hours → possibly high marks
    - Low attendance + high study_hours + easy exam → likely high marks
    - High sleep_hours + high study_hours → good marks

Correlation with Exam Score

Feature	Correlation	Strength
Study Hours	0.718	Strong
Class Attendance	0.309	Moderate
Sleep Hours	0.133	Weak

Interpretation: study_hours is the most predictive; attendance has moderate influence; sleep hours alone has weak effect.

Linear Regression

Steps:

1. Drop student_id

2. Encode categorical features
3. Split into training (80%) and testing (20%) sets
4. Train Linear Regression model
5. Evaluate using RMSE

Results:

- Validation RMSE ≈ 9.77
- Interpretation:
- On average, predictions are off by ±10 marks (reasonable for 0–100 range, but not strong)

Limitations of Linear Regression:

- Cannot model non-linear relationships (e.g., optimal sleep hours, difficulty impact)
- Cannot capture interactions (StudyHours × Attendance × Method)
- Treats categorical features as additive
- RMSE ~10 is expected for this dataset

Random Forest Regression

- Captures non-linear patterns and feature interactions
- Train RMSE ≈ 4.67, Validation RMSE ≈ 10.55 → overfitting observed
- Feature Importance (top 10):
    1. study_hours: 58.5%
    2. class_attendance: 15.9%
    3. sleep_hours: 6.5%
    4. age: 2.5%
    5. sleep_quality_poor: 2.3%
    6. facility_rating_low: 2.2%
    7. sleep_quality_good: 1.96%

8. study_method_self-study: 1.09%
9. study_method_online videos: 0.93%
10. facility_rating_medium: 0.77%

Reason for overfitting:

- Bounded exam scores (0–100)
- Noise from human behavior (sleep, study habits)
- Deep trees (max_depth=15) overfit sparse patterns

CatBoost Regressor

- Handles categorical features natively
- Train RMSE ≈ 9.47, Validation RMSE ≈ 9.78 → minimal overfitting
- Feature importance (CatBoost):
  - study_hours: 50.7%
  - class_attendance: 18.7%
  - sleep_quality: 9.6%
  - study_method: 8.3%
  - facility_rating: 6.8%
  - sleep_hours: 4.7%
  - Others <1%
- Slight hyperparameter adjustments → RMSE ≈ 9.82

Ensemble Model

- Combining predictions slightly improved RMSE
- Ensemble RMSE ≈ 9.77

Key Insights

1. **Most important features:** study_hours, class_attendance, sleep_quality, study_method
2. **Linear models** are limited due to non-linear patterns and interactions
3. **Random forests** may overfit without careful tuning
4. **CatBoost** is effective for categorical-heavy datasets and generalizes well
5. Noise and bounded score ranges limit the achievable RMSE (~9–10)