IF5100 - Pemrograman untuk Data Analitik

Tugas Besar
Kelompok 07
Analisis Data Exam Score Prediction

Bryan P. Hutagalung
18222130
18222130@std.stei.itb.ac.id

Andang Kurniawan
23524061
23524061@std.stei.itb.ac.id

Silvia Rahma
23525021
23525021@std.stei.itb.ac.id

Table of contents

Import Dataset and Libraries

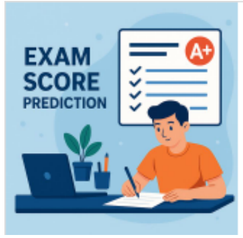
1

```
Shape: (20000, 13)
```

	student_id	age	gender	course	study_hours	class_attendance	internet_access	sleep_hours	sleep_quality	study_method	facility_rating	exam_difficulty	exam_score
0	1	17	male	diploma	2.78	92.9	yes	7.4	poor	coaching	low	hard	58.9
1	2	23	other	bca	3.37	64.8	yes	4.6	average	online videos	medium	moderate	54.8
2	3	22	male	b.sc	7.88	76.8	yes	8.5	poor	coaching	high	moderate	90.3
3	4	20	other	diploma	0.67	48.4	yes	5.8	average	online videos	low	moderate	29.7
4	5	20	female	diploma	0.89	71.6	yes	9.8	poor	coaching	low	moderate	43.7

```
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   student_id         20000 non-null  int64
 1   age                20000 non-null  int64
 2   gender             20000 non-null  object
 3   course             20000 non-null  object
 4   study_hours        20000 non-null  float64
 5   class_attendance   20000 non-null  float64
 6   internet_access    20000 non-null  object
 7   sleep_hours        20000 non-null  float64
 8   sleep_quality      20000 non-null  object
 9   study_method       20000 non-null  object
 10  facility_rating    20000 non-null  object
 11  exam_difficulty    20000 non-null  object
 12  exam_score         20000 non-null  float64
dtypes: float64(4), int64(2), object(7)
memory usage: 2.0+ MB
Missing values: 0
Duplicates: 0
```

- **Total 20.000 baris dan 13 fitur**
- **Tidak ada missing values dan tidak ada duplikasi**
- **Fitur mencakup jam belajar, kehadiran, akses internet, tidur, metode belajar, fasilitas, dan nilai ujian**
- **Variabel target: exam_score**
- **Dataset relatif seimbang dan representatif untuk modelling prediktif**
- **Preprocessing fokus pada encoding variabel kategorikal dan outlier removal, karena dataset sudah bersih**

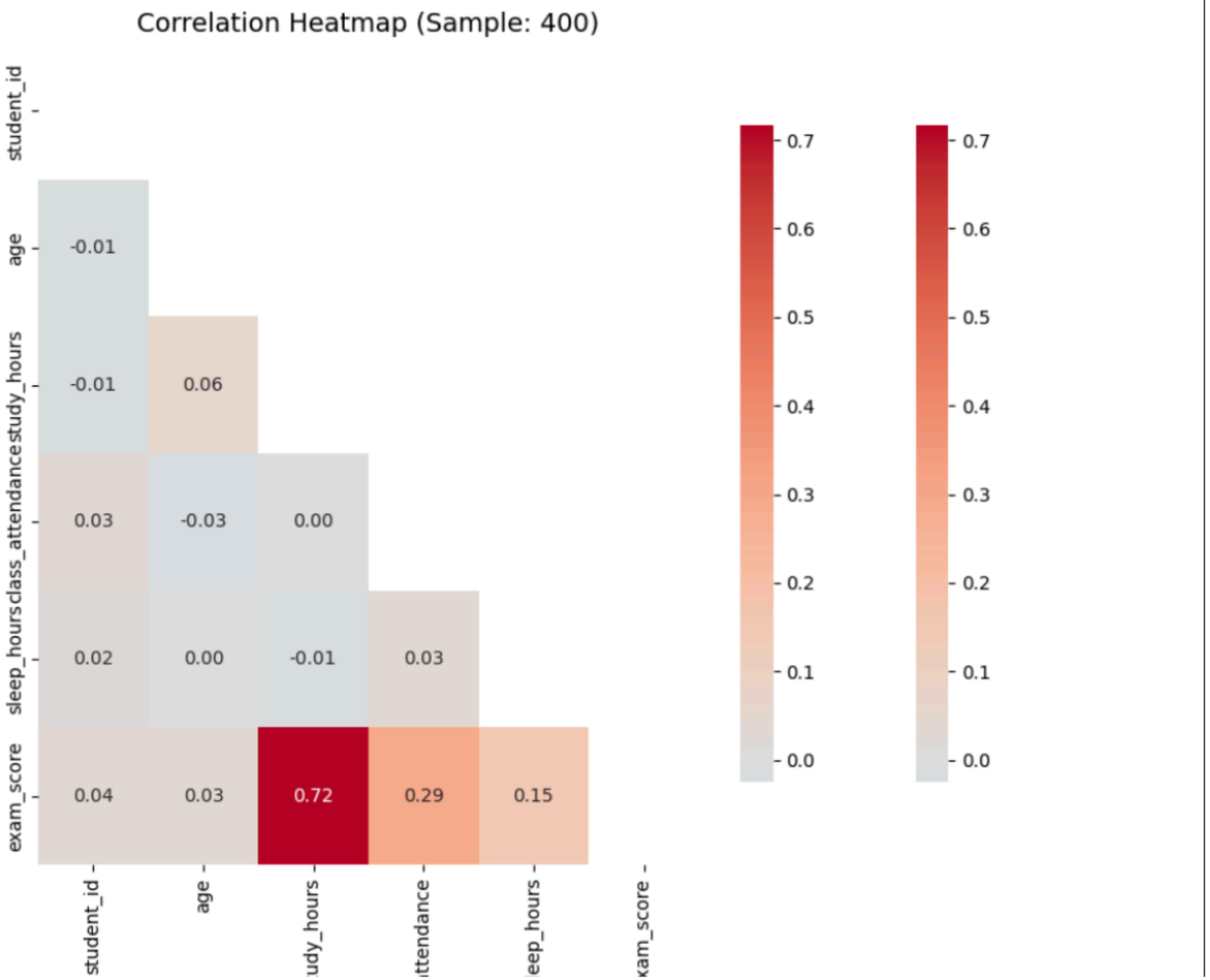


Exam Score Prediction Dataset

A detailed dataset containing academic, behavioral, lifestyle, and environmental

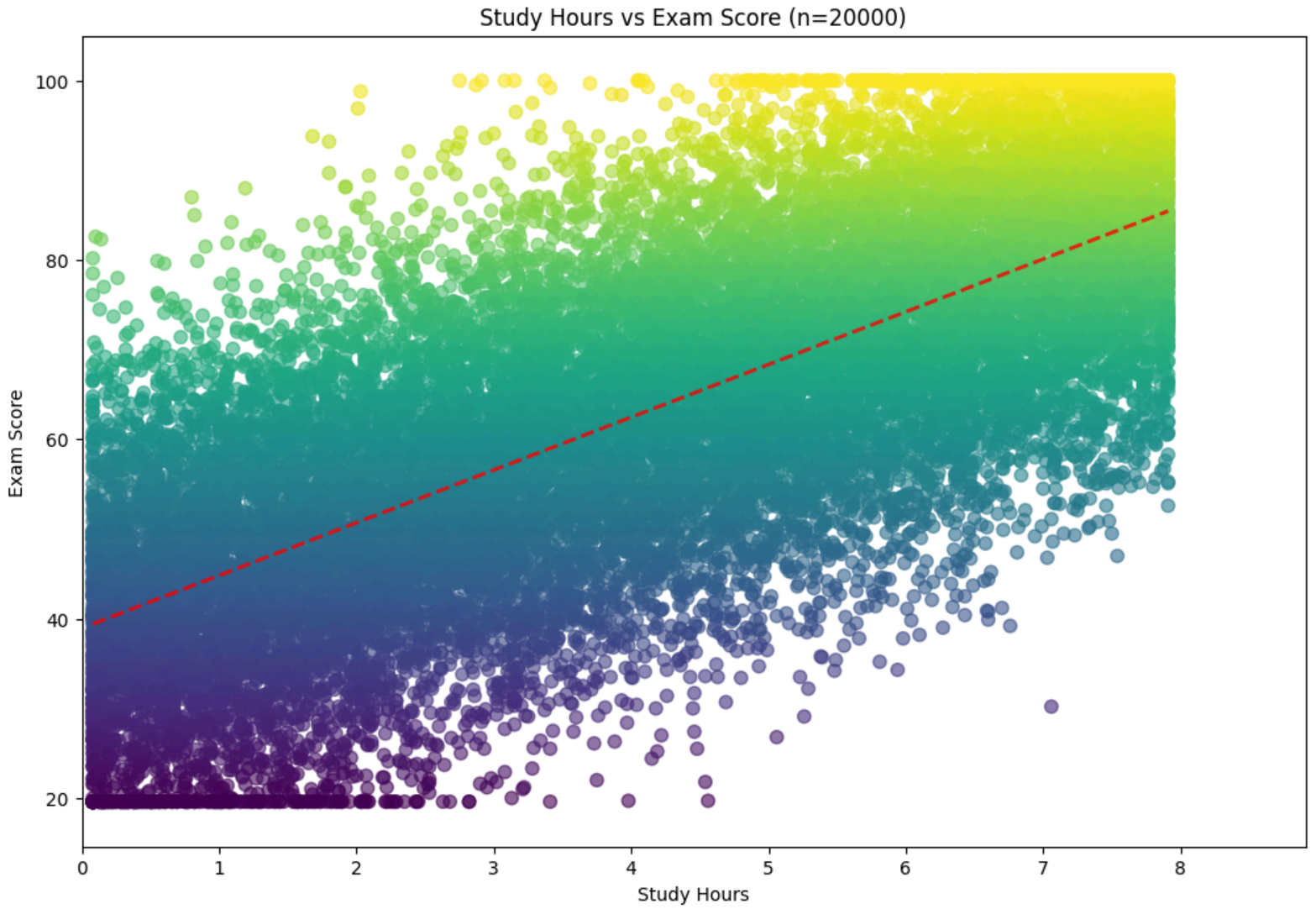
kaggle.com

Correlation Heatmap



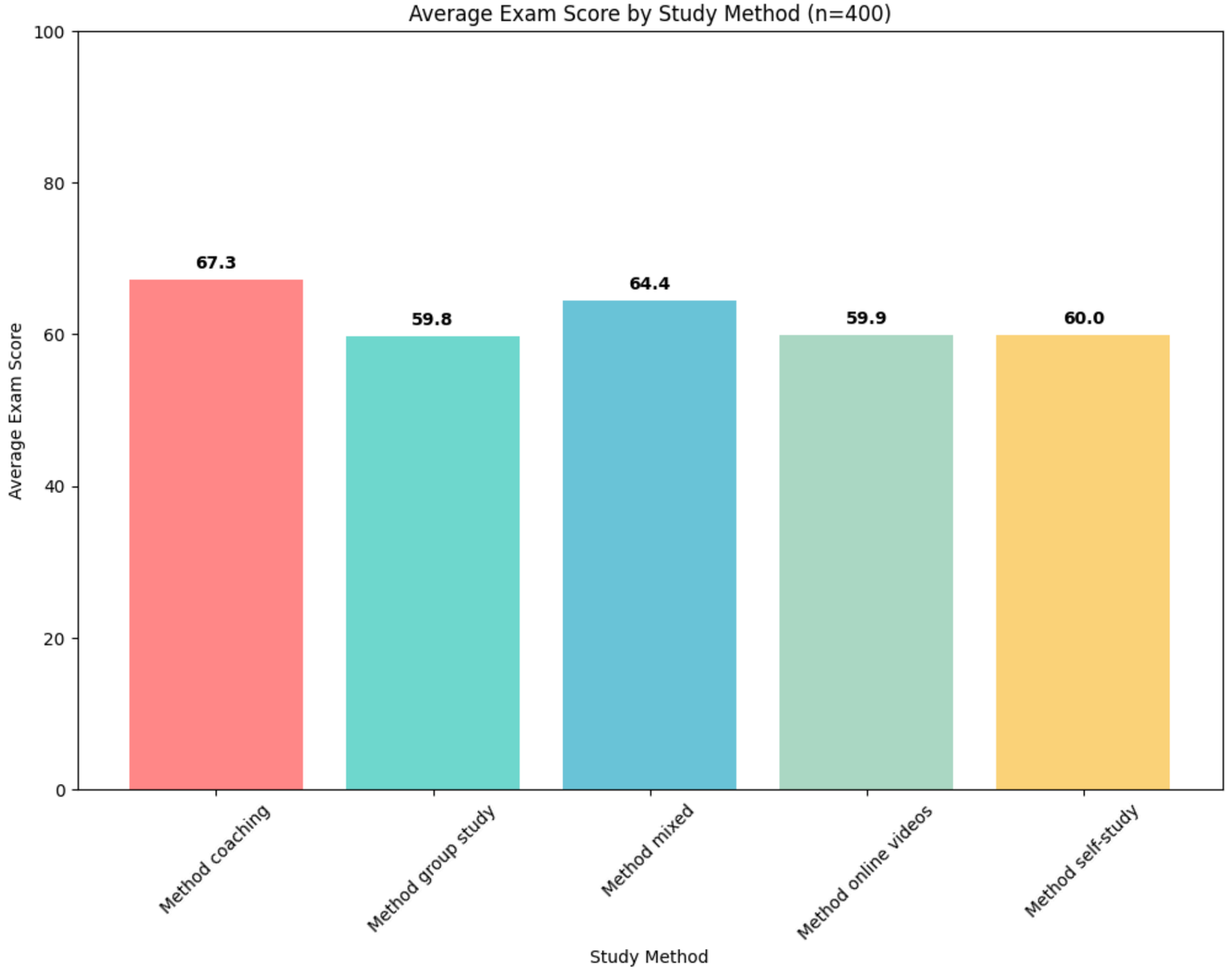
- Study_hours memiliki korelasi tertinggi dengan exam_score (~0.72).
- Class_attendance dan sleep_hours menunjukkan korelasi positif moderat.
- Pola korelasi stabil pada berbagai ukuran sampel (divalidasi dengan animasi).
- Tidak ada multikolinearitas ekstrem di fitur utama.

Study Hours vs Exam Score



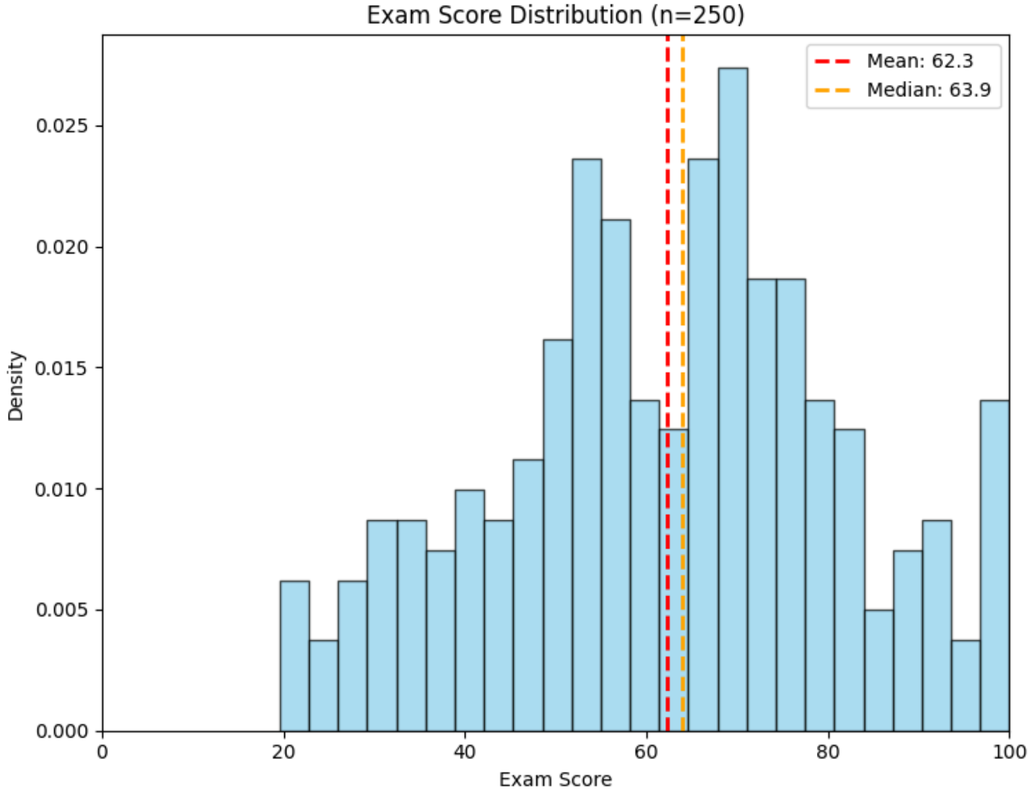
- Scatter plot menunjukkan hubungan positif antara jam belajar dan nilai ujian.
- Siswa dengan jam belajar lebih tinggi cenderung memperoleh nilai yang lebih baik.
- Garis tren memperlihatkan pola korelasi yang konsisten pada berbagai ukuran sampel.

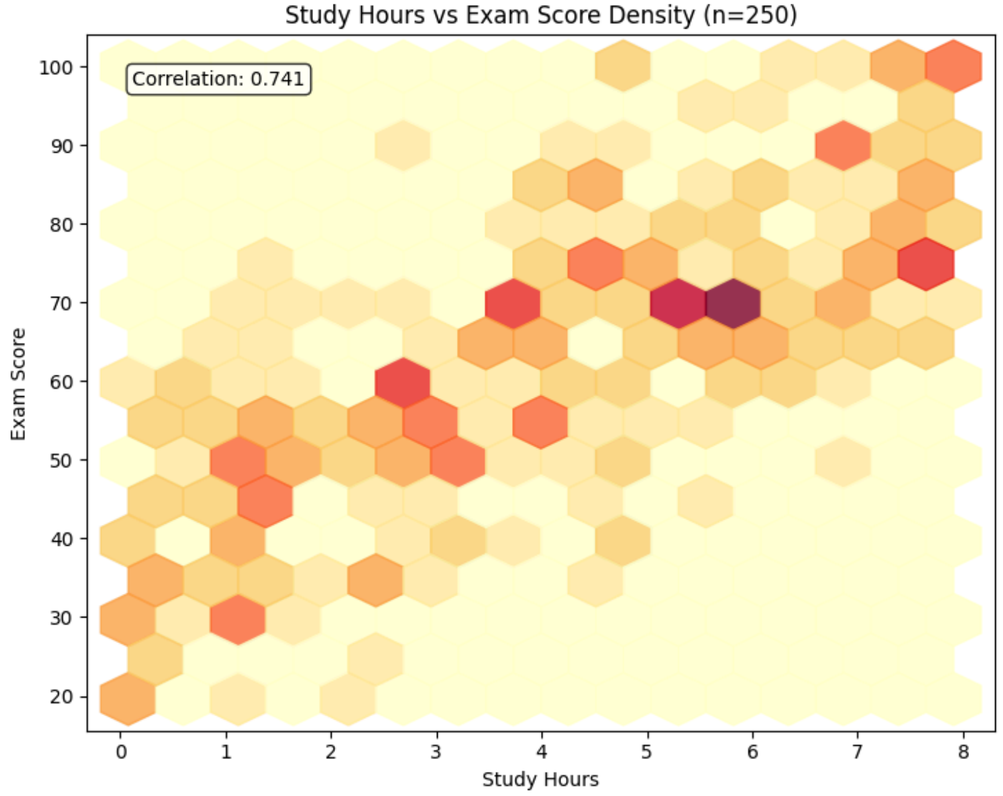
Study Method Performance



- Metode coaching menghasilkan rata-rata nilai tertinggi.
- Online videos dan self-study menunjukkan performa menengah yang stabil.
- Pola perbedaan performa antar metode konsisten pada berbagai ukuran sampel.

Distribution Evolution





- **Distribusi nilai ujian terlihat stabil dan mendekati pola normal.**
- **Korelasi antara jam belajar dan nilai ujian tetap kuat saat ukuran sampel diperbesar.**
- **Hexbin plot memperlihatkan area konsentrasi data yang menunjukkan pola hubungan yang konsisten.**

Data Cleaning

01

Cek White Space

Cek :
- white space di awal
- white space di akhir
- multi white space

Result :
- white space di awal teks : 0
- white space di akhir teks : 0
- Multi white space antar kata : 0

02

Cek Missing Value

Result : Tidak ada missing value
- student_id 0
- age 0
- gender 0
- course 0
- study_hours 0
- class_attendance 0
- internet_access 0
- sleep_hours 0
- sleep_quality 0
- study_method 0
- facility_rating 0
- exam_difficulty 0
- exam_score 0

03

Cek Data Duplicate

Result : Tidak ada data duplicated

04

Cek Outlier

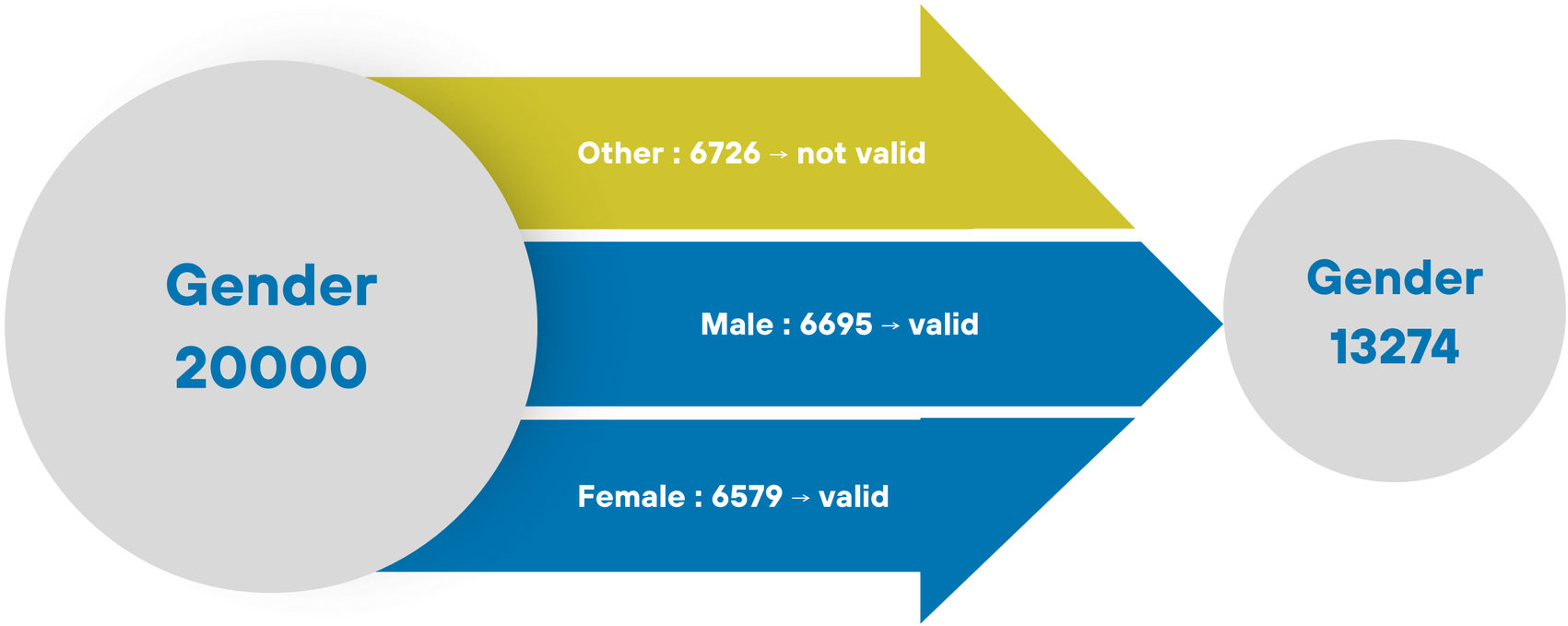
Cek : formula IQR (Interquartile Range)
- Q1 = nilai 25% data terurut
- Q3 = nilai 75% data terurut
- lower = Q1 - 1.5 * (Q3 - Q1)
- upper = Q3 + 1.5 * (Q3 - Q1)
- Outlier → data < lower | data > upper

Result : Tidak ada data outlier
- Outlier student_id : 0
- Outlier age : 0
- Outlier study_hours : 0
- Outlier class_attendance : 0
- Outlier sleep_hours: 0
- Oulier exam_score : 0

Data Validation

Validation : membersihkan data yang tidak valid berdasarkan aturan logis

Atribut gender : tidak valid → ‘other’



Feature Engineering & Encoding

01

Feature Creation

Fitur Baru :

- **study_efficiency: exam_score / study_hours**
  Menangkap hubungan seberapa efektif waktu belajar seseorang dalam menghasilkan nilai ujian

- **sleep_study_ratio: sleep_hours / study_hours**
  Menggambarkan keseimbangan antara waktu tidur dan waktu belajar.

02

Feature Encoding

Label Encoder :

- Gender
- Course
- internet_access
- sleep_quality
- study_method
- facility_rating
- exam_difficulty

	student_id	age	gender	course	study_hours	class_attendance	internet_access	sleep_hours	sleep_quality	study_method	facility_rating	exam_difficulty	exam_score	study_efficiency	sleep_study_ratio
0	1	17	1	6	2.78	92.9	1	7.4	2	0	1	1	58.9	21.187050	2.661871
2	3	22	1	1	7.88	76.8	1	8.5	2	0	0	2	90.3	11.459391	1.078680
4	5	20	0	6	0.89	71.6	1	9.8	2	0	1	2	43.7	49.101124	11.011236
5	6	23	1	2	3.48	65.4	1	4.2	1	2	1	2	58.2	16.724138	1.206897
6	7	17	0	2	1.35	69.0	1	7.4	0	3	0	1	53.7	39.777778	5.481481

Data Overview

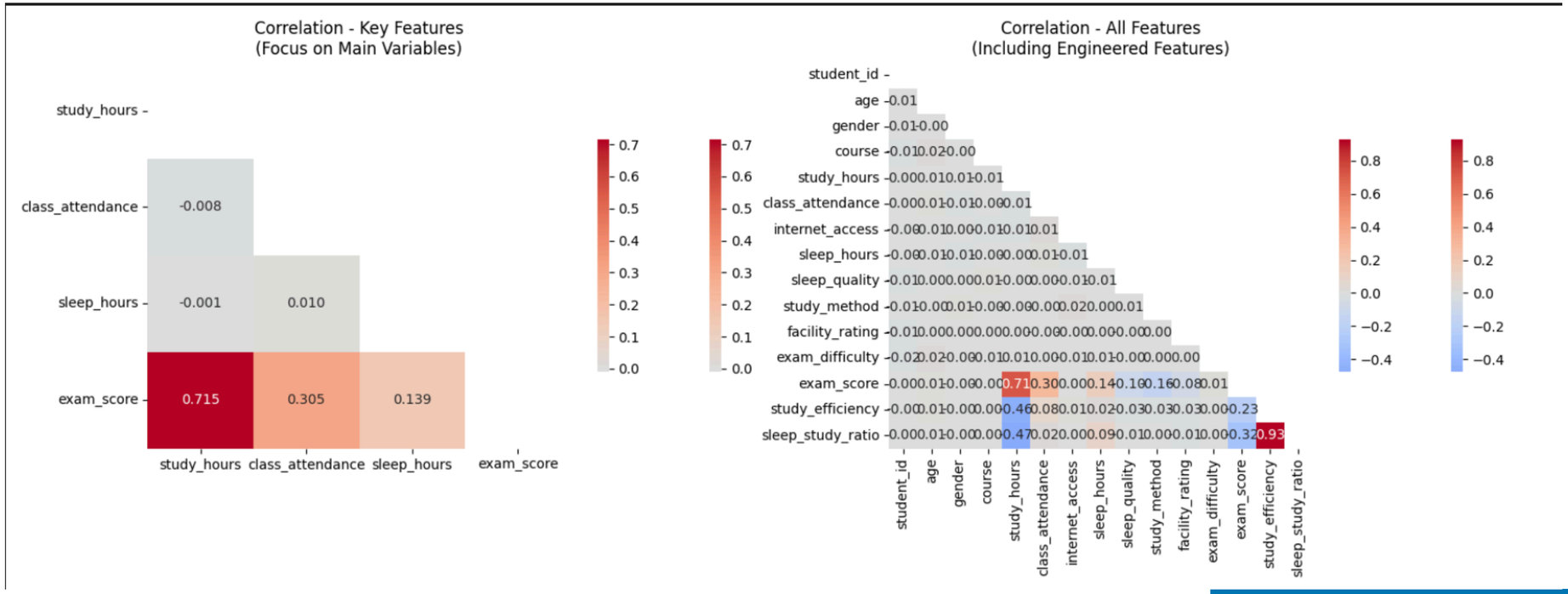

```
Dataset setelah preprocessing:
Shape: (13274, 15)
Total fitur: 15
Fitur numerik: 15
Fitur kategorikal (telah di-encode): 0

Statistik deskriptif fitur utama:
       study_hours  class_attendance  sleep_hours  exam_score  \
count     13274.00          13274.00     13274.00    13274.00
mean          4.01             70.00         7.01       62.49
std           2.30             17.34         1.73       18.87
min           0.08             40.60         4.10       19.60
25%           2.01             55.00         5.50       48.90
50%           4.05             69.90         7.00       62.60
75%           5.99             85.00         8.50       76.10
max           7.91             99.40         9.90      100.00

       study_efficiency  sleep_study_ratio
count          13274.00           13274.00
mean              33.28               4.86
std               67.63              11.46
min                4.28               0.52
```

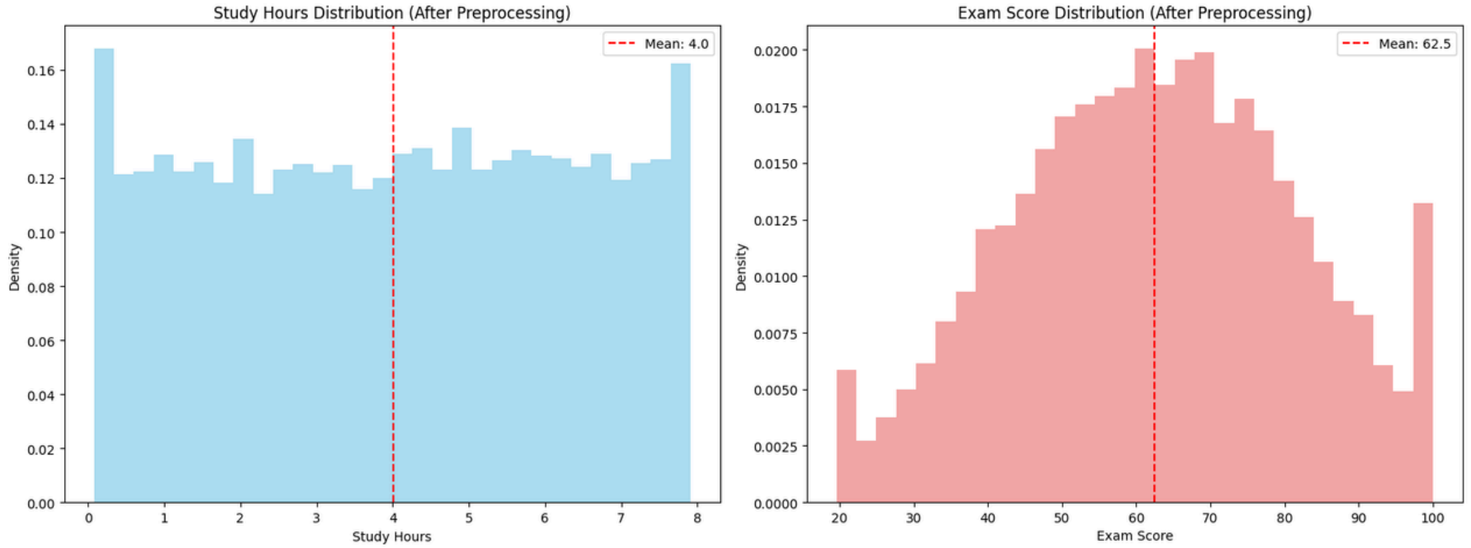
- Setelah preprocessing: 13.274 observasi, 15 fitur numerik (fitur kategorikal sudah di-encode).
- Fitur utama yang dianalisis: study_hours, class_attendance, sleep_hours, dan exam_score.
- Fitur turunan:
  - study_efficiency = kombinasi jam belajar & nilai ujian
  - sleep_study_ratio = keseimbangan waktu tidur vs belajar
- Terlihat variasi yang cukup besar dan beberapa outlier, sehingga perlu penanganan sebelum pemodelan.

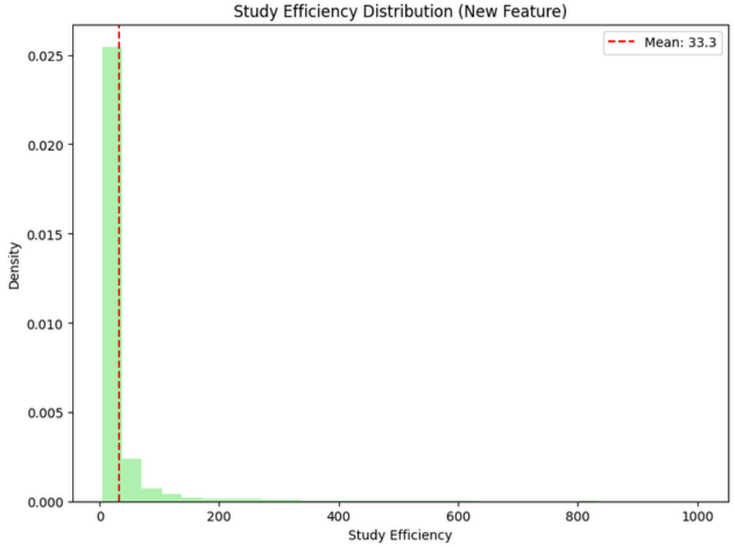
Correlation Matrix Comparison

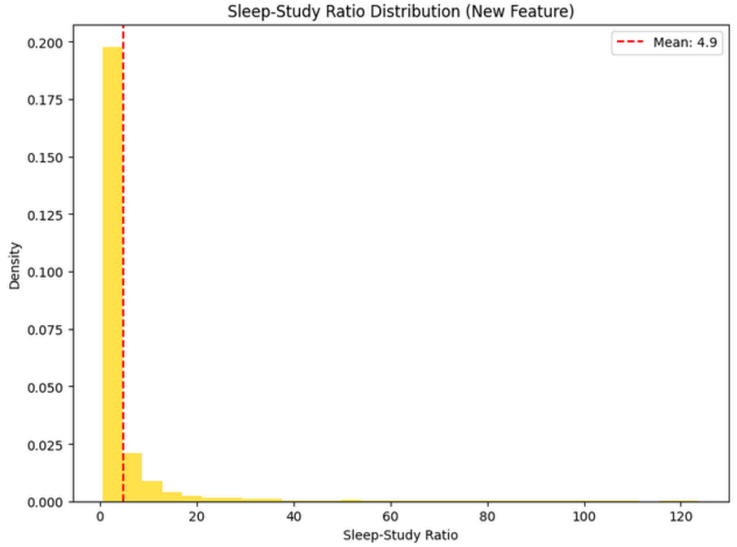


- **study_hours memiliki korelasi terkuat dengan exam_score (≈ 0,71).**
- **Fitur turunan (study_efficiency & sleep_study_ratio) menambah pola baru, termasuk korelasi negatif yang tidak muncul pada fitur asli.**
- **Setelah outlier removal, korelasi antar variabel menjadi lebih stabil dan mudah diinterpretasi.**
- **Variabel lain seperti class_attendance dan exam_difficulty tetap berpengaruh, tetapi dengan korelasi lebih kecil.**

Feature Distribution Analysis

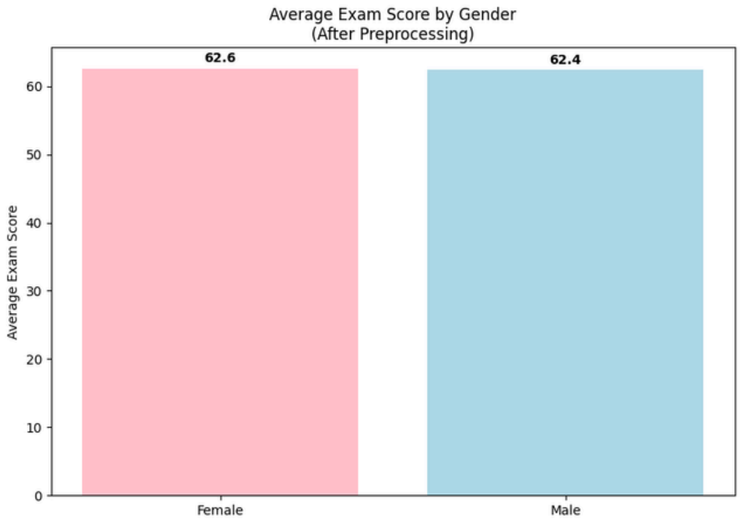


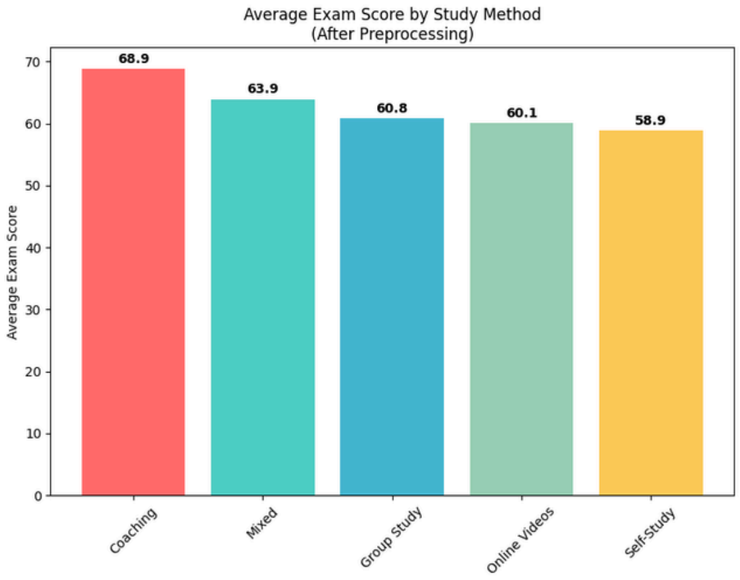


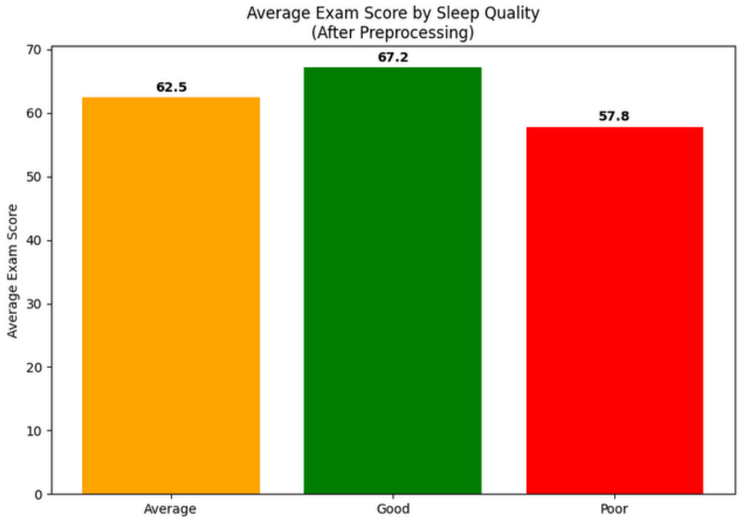


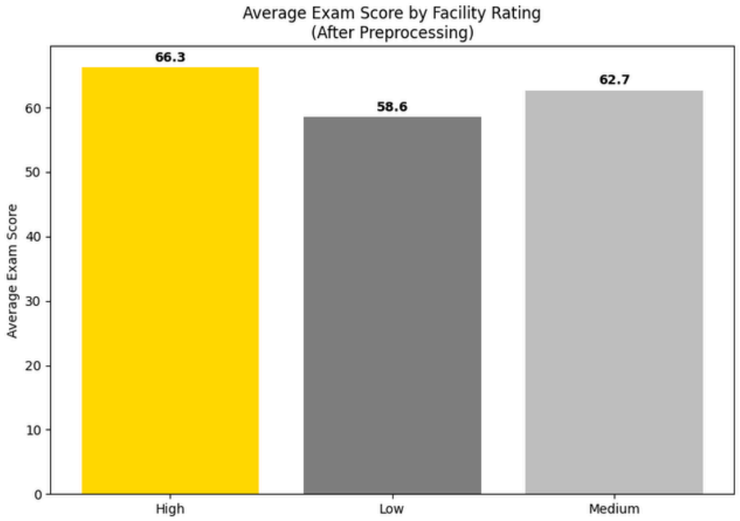
Analisis distribusi setelah preprocessing:
1. **Study hours: Distribusi lebih normal setelah outlier removal**
2. **Exam score: Range nilai lebih fokus, outlier ekstrem telah dihilangkan**
3. **Study efficiency: Feature baru menunjukkan distribusi yang informatif**
4. **Sleep-study ratio: Memberikan insight baru tentang balance siswa**

Categorical Analysis After Pre-Processing



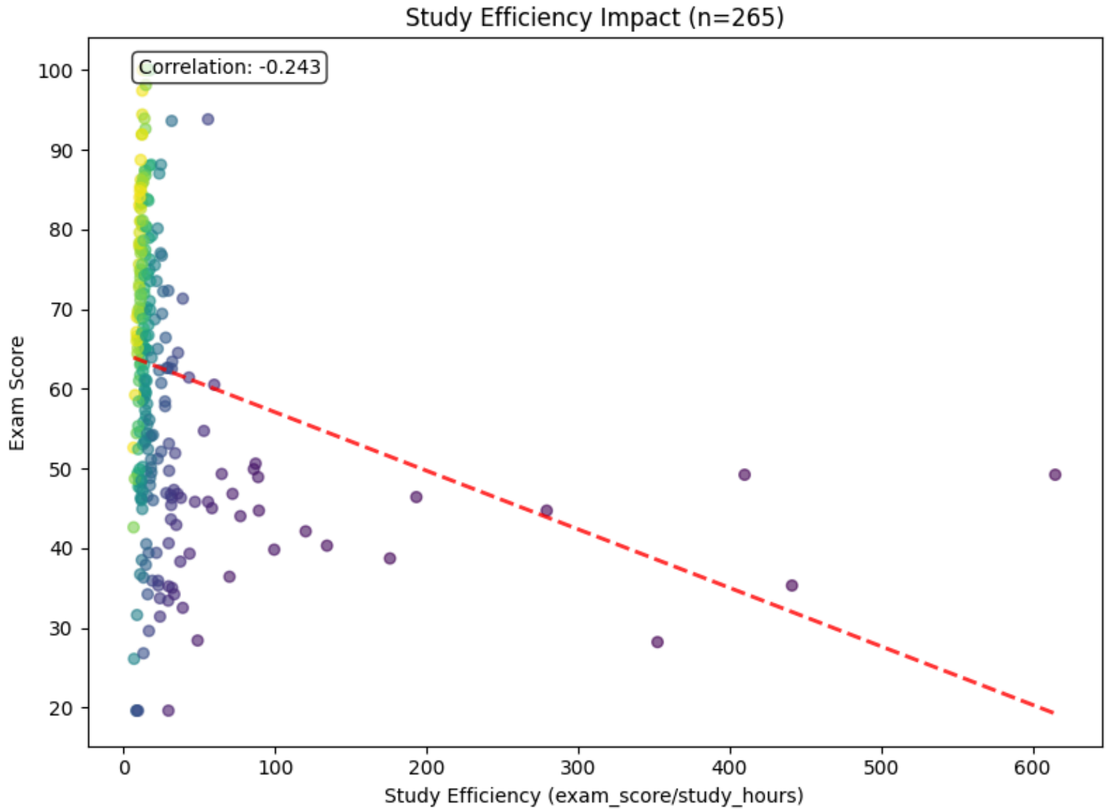


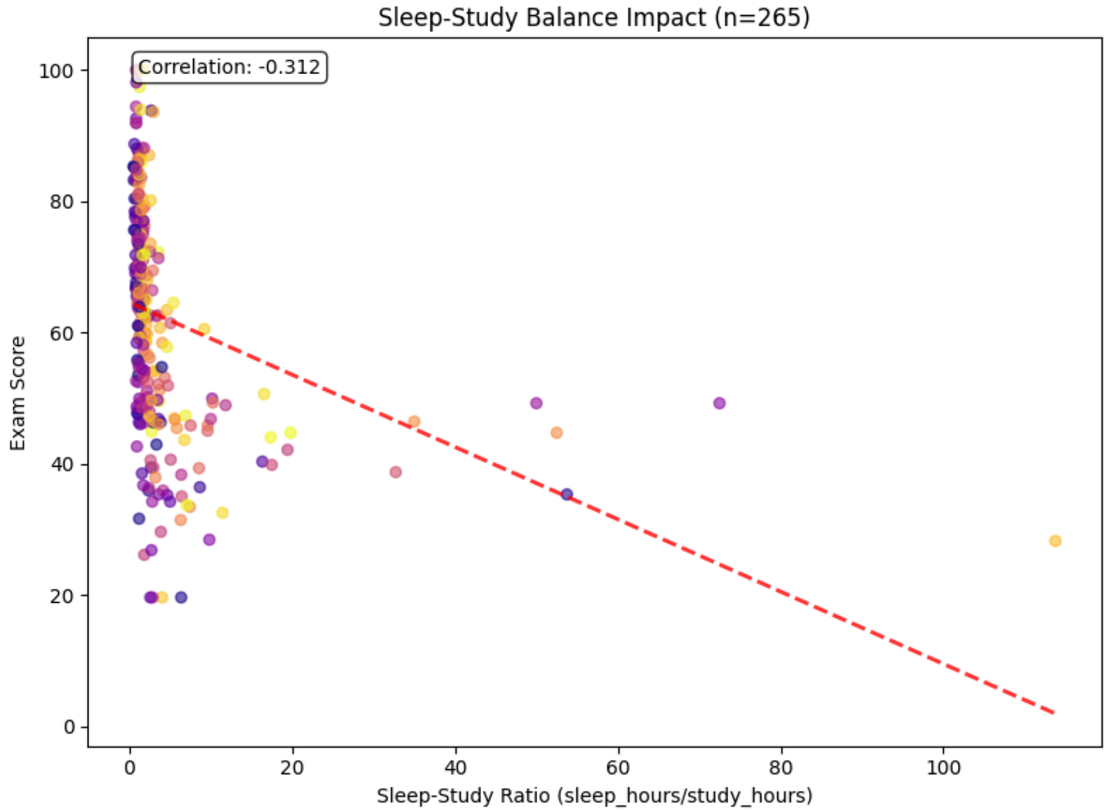




- **Gender: Performa relatif sama; pola tetap stabil setelah cleaning.**
- **Study Method: Coaching memberikan skor tertinggi secara konsisten.**
- **Sleep Quality: Setelah outlier removal, hubungan positif menjadi lebih jelas.**
- **Facility Rating: Lingkungan belajar yang lebih baik tercermin pada skor ujian lebih tinggi.**

Feature Engineering Impact Analysis





• Dua fitur baru (study_efficiency dan sleep_study_ratio) memberikan perspektif baru tentang efektivitas belajar dan keseimbangan tidur belajar. Keduanya menunjukkan korelasi negatif terhadap nilai ujian (≈ -0,24 dan ≈ -0,31), sehingga membantu model menangkap pola perilaku belajar yang lebih kompleks.

Feature Selection & Train Test Split

Selected Features

1. 'study_hours'
2. 'sleep_study_ratio'
3. 'class_attendance'
4. 'study_efficiency'
5. 'study_method'
6. 'sleep_hours'
7. 'sleep_quality'

Feature Corrleation with Exam Score (Abs)

1. study_hours: 0.714765
2. sleep_study_ratio: 0.317278
3. class_attendance: 0.304570
4. study_efficiency: 0.225766
5. study_method: 0.156779
6. sleep_hours: 0.139298
7. sleep_quality: 0.100711

Train Test Split

- Training set: (10619, 7)
- Test set: (2655, 7)

Model Evaluation & Selection

Linear Regression

- R2 Score: 0.689
- RMSE: 10.46
- MAE: 8.29
- CV R2: 0.692 +/- 0.008

Random Forest

- R2 Score: 0.996
- RMSE: 1.16
- MAE: 0.51
- CV R2: 0.995 +/- 0.001

Ridge

- R2 Score: 0.689
- RMSE: 10.46
- MAE: 8.29
- CV R2: 0.692 +/- 0.008

Gradient Boosting

- R2 Score: 0.964
- RMSE: 3.54
- MAE: 2.33
- CV R2: 0.965 +/- 0.003

Random Forest

Model Justification, Application

Alasan Pembuatan Model

1. **Tujuan**: Memprediksi nilai ujian siswa berdasarkan pola belajar
2. **Problem type**: Regression (target variabel kontinyu)
3. **Dataset**: 13,274 siswa dengan 13 variabel independen

Potensi Penggunaan di Dunia Nyata

1. Sistem rekomendasi jam belajar optimal untuk siswa
2. Identifikasi siswa berisiko rendah nilai ujian
3. Personalisasi strategi pembelajaran berdasarkan profil siswa
4. Optimasi alokasi sumber daya pendidikan
5. Konsultasi akademik berbasis data

Alasan Pemilihan Atribut

Atribut dipilih berdasarkan korelasi dengan exam_score:

1. study_hours: korelasi = 0.715
2. sleep_study_ratio: korelasi = -0.317
3. class_attendance: korelasi = 0.305
4. study_efficiency: korelasi = -0.226
5. study_method: korelasi = -0.157
6. sleep_hours: korelasi = 0.139
7. sleep_quality: korelasi = -0.101

Feature Importance (Random Forest)

1. study_hours: 0.671
2. study_efficiency: 0.301
3. class_attendance: 0.023
4. sleep_study_ratio: 0.002
5. sleep_hours: 0.002
6. study_method: 0.001
7. sleep_quality: 0.001

Data Insights

Dataset Analysis

- 13274 siswa dianalisis dari 20000 data awal
- 7 fitur terpilih dari 15 total fitur

Model Performance

- Best model: Random Forest
- Akurasi prediksi: R2 = 99.6%
- Error rate: RMSE = 1.2 poin
- Mean error: MAE = 0.5 poin

Key Findings

- Study hours adalah prediktor terkuat (importance: 0.671)
- Model mampu menjelaskan 99.6% variabilitas nilai ujian
- Fitur engineered (study_efficiency) berkontribusi signifikan

Actionable Insights

- Siswa disarankan belajar minimal 4-6 jam per hari
- Metode coaching memberikan hasil terbaik
- Balance antara jam tidur dan belajar penting untuk performa optimal

“ITB”

“Iqro, Tabayyun, Bertindak”

- Prof. Dr. Ir. Suhono H. Supangkat -

Thank You

Bryan P. Hutagalung
18222130
18222130@std.stei.itb.ac.id

Andang Kurniawan
23524061
23524061@std.stei.itb.ac.id

Silvia Rahma
23525021
23525021@std.stei.itb.ac.id