Задача классификации

Логистическая регрессия

Подготовил: студент группы М8О-307Б-23
Бельский Г. Б.

Почему логистическая регрессия?

Подходит для объяснимых и реальных задач: чётко показывает вклад каждого признака, легко интерпретируется и устойчива к мультиколлинеарности, особенно с регуляризацией.

Прекрасно работает на инженерных и табличных данных со смешанными признаками — как в моем датасете

Легко масштабируется и прост в настройке

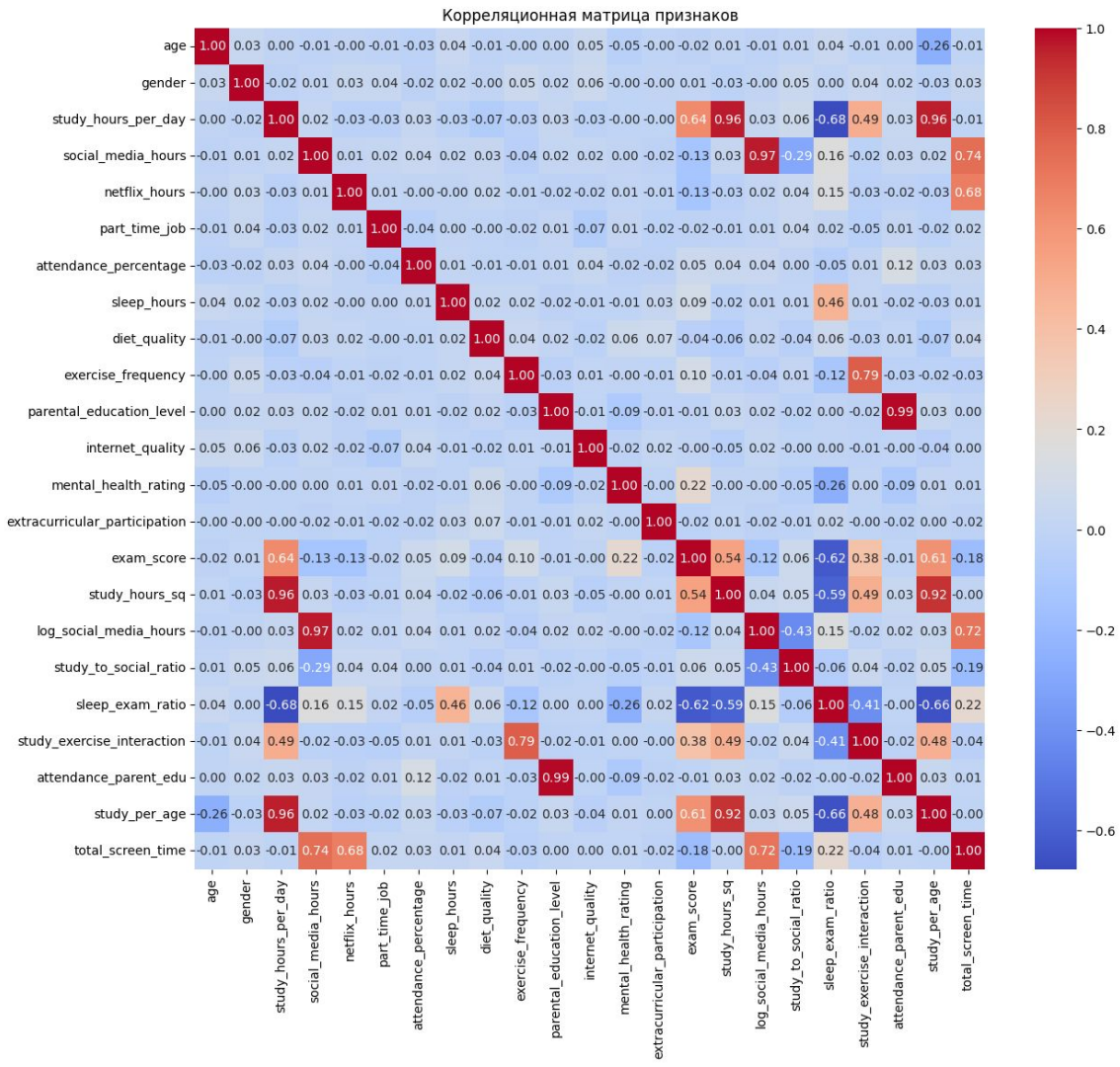
Подготовка датасета

Очистка: удаляем идентификаторы, парсим категориальные столбцы через LabelEncoder, числовые используем напрямую.

Feature Engineering: вы создали кучу новых фич — квадраты, логарифмы, отношения и произведения. Это усиливает модель и покрывает нелинейные зависимости.

Удаление выбросов: через IQR для корректности метрик.

Таргет: target_class = (exam_score >= 60).astype(int). Классы сбалансированы после разбивки.


Корреляционная матрица признаков
age 1.00 0.03 0.00 -0.01 -0.00 -0.01 -0.03 0.04 -0.01 -0.00 0.00 0.05 -0.05 -0.00 -0.02 0.01 -0.01 0.01 0.04 -0.01 0.00 -0.26 -0.01
gender 0.03 1.00 -0.02 0.01 0.03 0.04 -0.02 0.02 -0.00 0.05 0.02 0.06 -0.00 -0.00 0.01 -0.03 -0.00 0.05 0.00 0.04 0.02 -0.03 0.03
study_hours_per_day 0.00 -0.02 1.00 0.02 -0.03 -0.03 0.03 -0.03 -0.07 -0.03 0.03 -0.03 -0.00 -0.00 0.64 0.96 0.03 0.06 -0.68 0.49 0.03 0.96 -0.01
social_media_hours -0.01 0.01 0.02 1.00 0.01 0.02 0.04 0.02 0.03 -0.04 0.02 0.02 0.00 -0.02 -0.13 0.03 0.97 -0.29 0.16 -0.02 0.03 0.02 0.74
netflix_hours -0.00 0.03 -0.03 0.01 1.00 0.01 -0.00 -0.00 0.02 -0.01 -0.02 -0.02 0.01 -0.01 -0.13 -0.03 0.02 0.04 0.15 -0.03 -0.02 -0.03 0.68
part_time_job -0.01 0.04 -0.03 0.02 0.01 1.00 -0.04 0.00 -0.00 -0.02 0.01 -0.07 0.01 -0.02 -0.02 -0.01 0.01 0.04 0.02 -0.05 0.01 -0.02 0.02
attendance_percentage -0.03 -0.02 0.03 0.04 -0.00 -0.04 1.00 0.01 -0.01 -0.01 0.01 0.04 -0.02 -0.02 0.05 0.04 0.04 0.00 -0.05 0.01 0.12 0.03 0.03
sleep_hours 0.04 0.02 -0.03 0.02 -0.00 0.00 0.01 1.00 0.02 0.02 -0.02 -0.01 -0.01 0.03 0.09 -0.02 0.01 0.01 0.46 0.01 -0.02 -0.03 0.01
diet_quality -0.01 -0.00 -0.07 0.03 0.02 -0.00 -0.01 0.02 1.00 0.04 0.02 -0.02 0.06 0.07 -0.04 -0.06 0.02 -0.04 0.06 -0.03 0.01 -0.07 0.04
exercise_frequency -0.00 0.05 -0.03 -0.04 -0.01 -0.02 -0.01 0.02 0.04 1.00 -0.03 0.01 -0.00 -0.01 0.10 -0.01 -0.04 0.01 -0.12 0.79 -0.03 -0.02 -0.03
parental_education_level 0.00 0.02 0.03 0.02 -0.02 0.01 0.01 -0.02 0.02 -0.03 1.00 -0.01 -0.09 -0.01 -0.01 0.03 0.02 -0.02 0.00 -0.02 0.99 0.03 0.00
internet_quality 0.05 0.06 -0.03 0.02 -0.02 -0.07 0.04 -0.01 -0.02 0.01 -0.01 1.00 -0.02 0.02 -0.00 -0.05 0.02 -0.00 0.00 -0.01 -0.00 -0.04 0.00
mental_health_rating -0.05 -0.00 -0.00 0.00 0.01 0.01 -0.02 -0.01 0.06 -0.00 -0.09 -0.02 1.00 -0.00 0.22 -0.00 -0.00 -0.05 -0.26 0.00 -0.09 0.01 0.01
extracurricular_participation -0.00 -0.00 -0.00 -0.02 -0.01 -0.02 -0.02 0.03 0.07 -0.01 -0.01 0.02 -0.00 1.00 -0.02 0.01 -0.02 -0.01 0.02 -0.00 -0.02 0.00 -0.02
exam_score -0.02 0.01 0.64 -0.13 -0.13 -0.02 0.05 0.09 -0.04 0.10 -0.01 -0.00 0.22 -0.02 1.00 0.54 -0.12 0.06 -0.62 0.38 -0.01 0.61 -0.18
study_hours_sq 0.01 -0.03 0.96 0.03 -0.03 -0.01 0.04 -0.02 -0.06 -0.01 0.03 -0.05 -0.00 0.01 0.54 1.00 0.04 0.05 -0.59 0.49 0.03 0.92 -0.00
log_social_media_hours -0.01 -0.00 0.03 0.97 0.02 0.01 0.04 0.01 0.02 -0.04 0.02 0.02 -0.00 -0.02 -0.12 0.04 1.00 -0.43 0.15 -0.02 0.02 0.03 0.72
study_to_social_ratio 0.01 0.05 0.06 -0.29 0.04 0.04 0.00 0.01 -0.04 0.01 -0.02 -0.00 -0.05 -0.01 0.06 0.05 -0.43 1.00 -0.06 0.04 -0.02 0.05 -0.19
sleep_exam_ratio 0.04 0.00 -0.68 0.16 0.15 0.02 -0.05 0.46 0.06 -0.12 0.00 0.00 -0.26 0.02 -0.62 -0.59 0.15 -0.06 1.00 -0.41 -0.00 -0.66 0.22
study_exercise_interaction -0.01 0.04 0.49 -0.02 -0.03 -0.05 0.01 0.01 -0.03 0.79 -0.02 -0.01 0.00 -0.00 0.38 0.49 -0.02 0.04 -0.41 1.00 -0.02 0.48 -0.04
attendance_parent_edu 0.00 0.02 0.03 0.03 -0.02 0.01 0.12 -0.02 0.01 -0.03 0.99 -0.00 -0.09 -0.02 -0.01 0.03 0.02 -0.02 -0.00 -0.02 1.00 0.03 0.01
study_per_age -0.26 -0.03 0.96 0.02 -0.03 -0.02 0.03 -0.03 -0.07 -0.02 0.03 -0.04 0.01 0.00 0.61 0.92 0.03 0.05 -0.66 0.48 0.03 1.00 -0.00
total_screen_time -0.01 0.03 -0.01 0.74 0.68 0.02 0.03 0.01 0.04 -0.03 0.00 0.00 0.01 -0.02 -0.18 -0.00 0.72 -0.19 0.22 -0.04 0.01 -0.00 1.00
age
gender
study_hours_per_day
social_media_hours
netflix_hours
part_time_job
attendance_percentage
sleep_hours
diet_quality
exercise_frequency
parental_education_level
internet_quality
mental_health_rating
extracurricular_participation
exam_score
study_hours_sq
log_social_media_hours
study_to_social_ratio
sleep_exam_ratio
study_exercise_interaction
attendance_parent_edu
study_per_age
total_screen_time
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Распределение данных

Классы разбиты почти поровну. Проведён предварительный анализ, boxplot и scatterplot-функции демонстрируют хорошую отделимость классов по основным переменным.

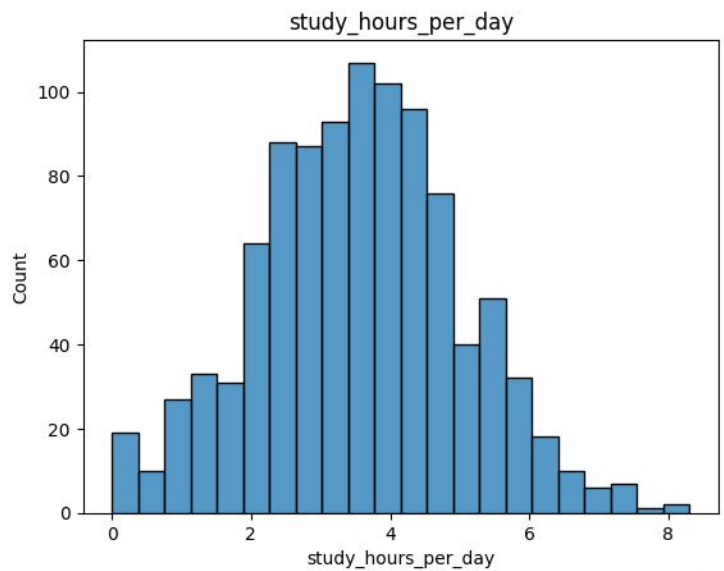
Корреляционная матрица позволяет обнаружить избыточные или зависимые признаки.

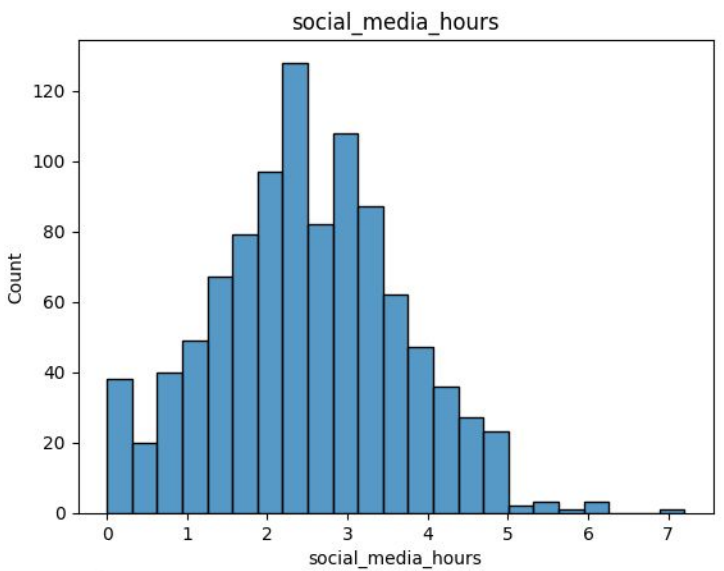
Результаты после удаления выбросов и логарифмических преобразований значительно улучшают компактность и нормальность распределений.

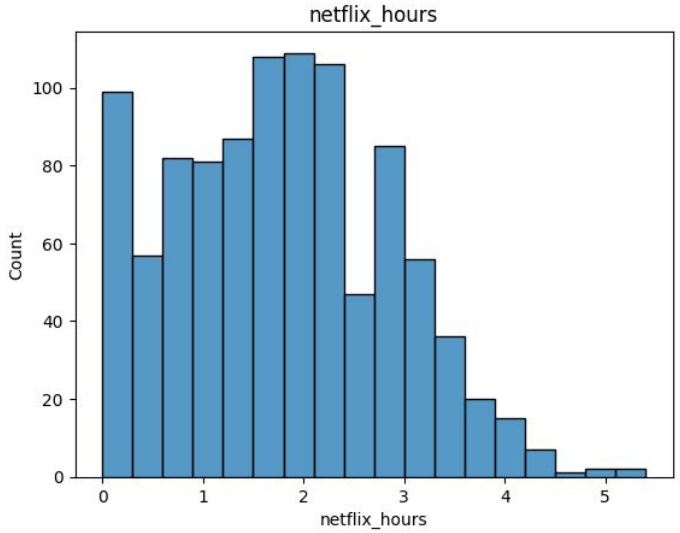
Настройки и метрики модели

Используем: LogisticRegression(penalty='l2', C=1.0, solver='lbfgs', max_iter=500, random_state=RANDOM_STATE) В Cross-Validation (StratifiedKFold, 5 фолдов):

Значение C	Accuracy (mean±std)	F1-score (mean±std)	ROC-AUC (mean±std)
0.01	~0.72 ± 0.03	~0.73 ± 0.03	~0.78 ± 0.03
0.1	~0.74 ± 0.02	~0.74 ± 0.03	~0.80 ± 0.02
1	~0.75 ± 0.02	~0.75 ± 0.02	~0.82 ± 0.02
10	~0.76 ± 0.02	~0.76 ± 0.02	~0.82 ± 0.01

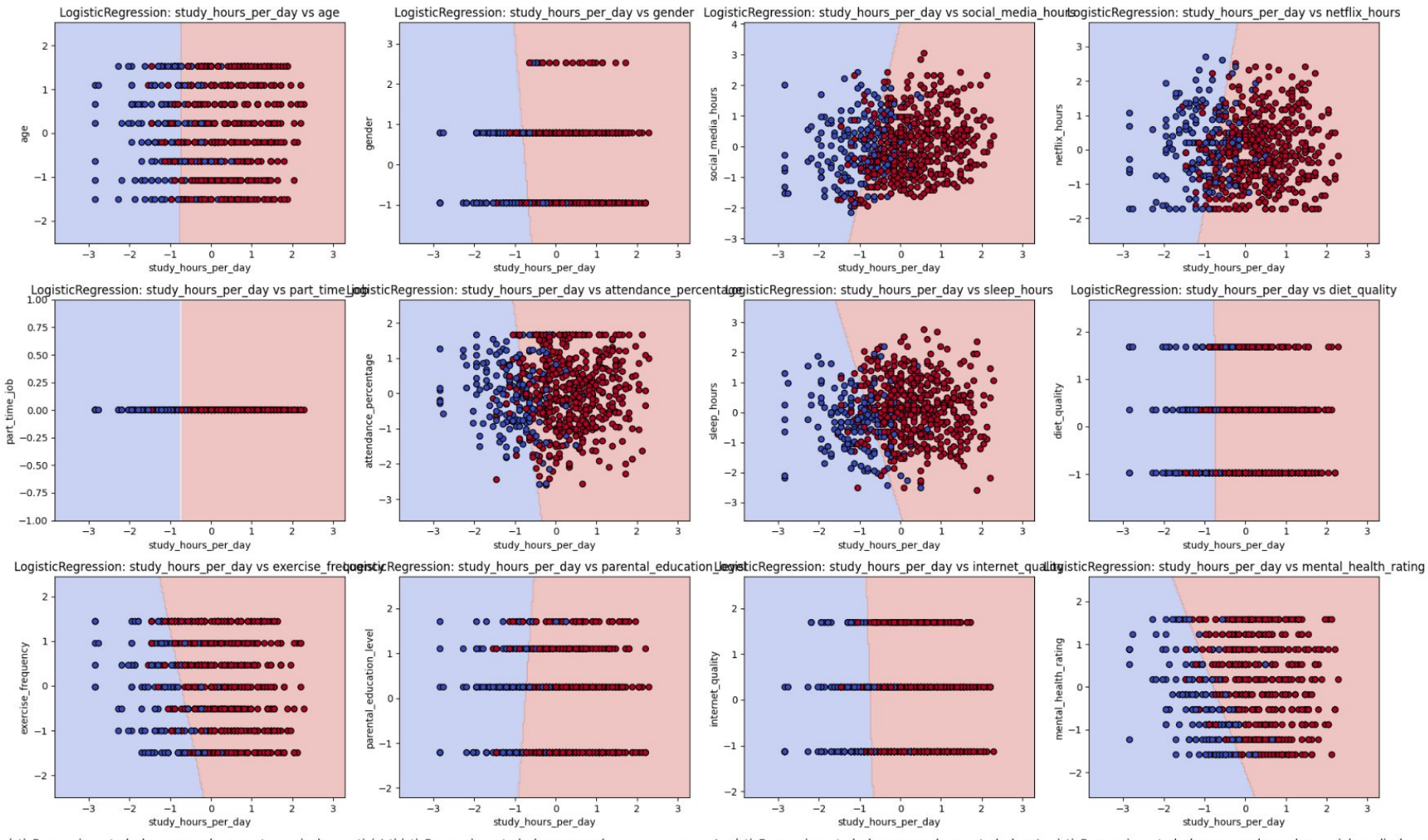

study_hours_per_day
Count
study_hours_per_day
0
20
40
60
80
100
0
2
4
6
8

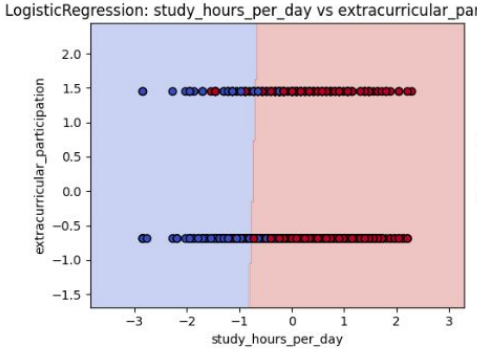

social_media_hours
Count
social_media_hours
0
20
40
60
80
100
120
0
1
2
3
4
5
6
7

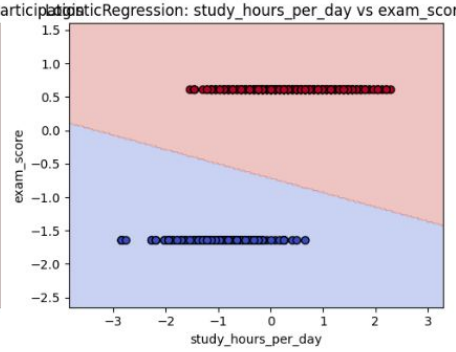

netflix_hours
Count
netflix_hours
0
20
40
60
80
100
0
1
2
3
4
5

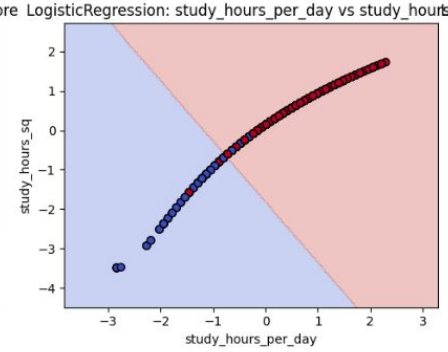
Визуализация (описательные графики)

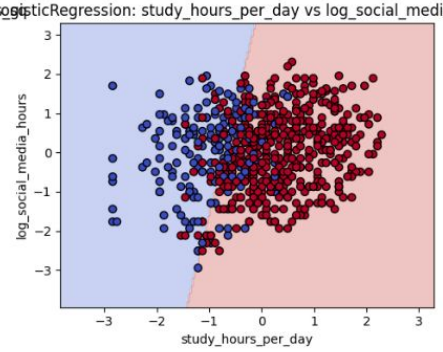
ROC-кривые (показывают, как модель разделяет классы; площадь под ROC >0.8 — очень хороший результат).

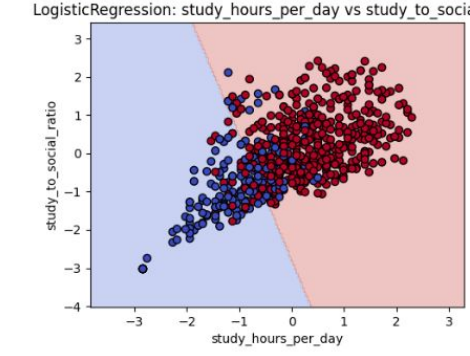

LogisticRegression: study_hours_per_day vs age
LogisticRegression: study_hours_per_day vs gender
LogisticRegression: study_hours_per_day vs social_media_hours
LogisticRegression: study_hours_per_day vs netflix_hours
LogisticRegression: study_hours_per_day vs part_time_job
LogisticRegression: study_hours_per_day vs attendance_percentage
LogisticRegression: study_hours_per_day vs sleep_hours
LogisticRegression: study_hours_per_day vs diet_quality
LogisticRegression: study_hours_per_day vs exercise_frequency
LogisticRegression: study_hours_per_day vs parental_education_level
LogisticRegression: study_hours_per_day vs internet_quality
LogisticRegression: study_hours_per_day vs mental_health_rating
study_hours_per_day
age
gender
social_media_hours
netflix_hours
part_time_job
attendance_percentage
sleep_hours
diet_quality
exercise_frequency
parental_education_level
internet_quality
mental_health_rating

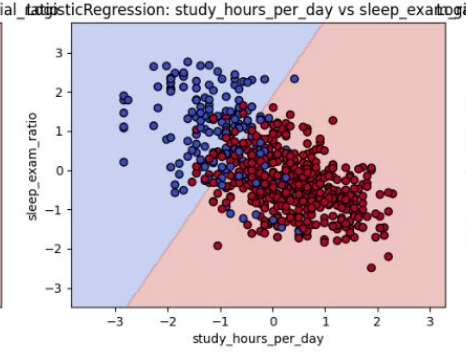

LogisticRegression: study_hours_per_day vs extracurricular_participation
extracurricular_participation
study_hours_per_day

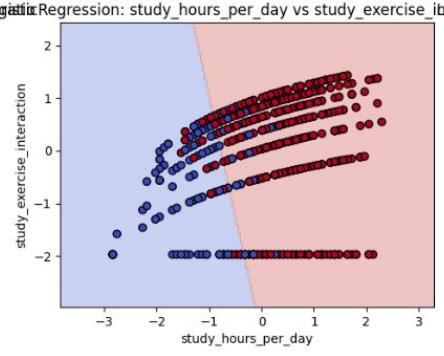

LogisticRegression: study_hours_per_day vs exam_score
exam_score
study_hours_per_day

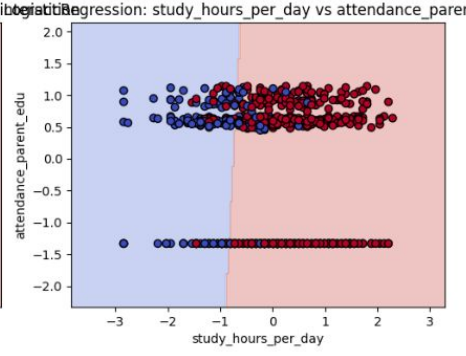

LogisticRegression: study_hours_per_day vs study_hours_sq
study_hours_sq
study_hours_per_day

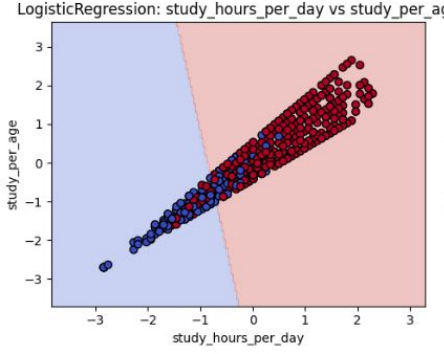

LogisticRegression: study_hours_per_day vs log_social_media_hours
log_social_media_hours
study_hours_per_day

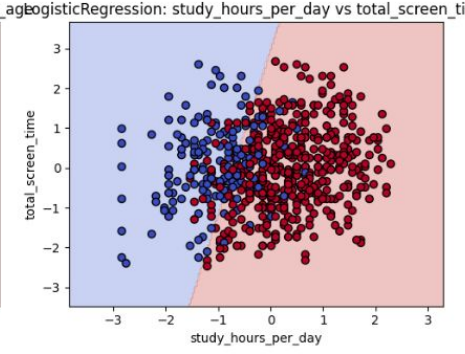

LogisticRegression: study_hours_per_day vs study_to_social_ratio
study_to_social_ratio
study_hours_per_day


LogisticRegression: study_hours_per_day vs sleep_exam_ratio
sleep_exam_ratio
study_hours_per_day


LogisticRegression: study_hours_per_day vs study_exercise_interaction
study_exercise_interaction
study_hours_per_day


LogisticRegression: study_hours_per_day vs attendance_parent_edu
attendance_parent_edu
study_hours_per_day


LogisticRegression: study_hours_per_day vs study_per_age
study_per_age
study_hours_per_day


LogisticRegression: study_hours_per_day vs total_screen_time
total_screen_time
study_hours_per_day

Почему выбирать LOGREG?

Лёгкая интерпретация результатов — важна для объяснимости (бизнес, медицина, HR-аналитика и большинство "продажных" задач).

Простая отладка, быстрая диагностика и контроль качества.

Можно быстро расширить: добавить нелинейностей, использовать как baseline для сравнения с более сложными ML/ensemble-моделями.

Полностью повторяемо и минимально зависит от тонких настроек.

Логистическая регрессия на этом датасете даст простой, быстрый и объяснимый старт для любой задачи бинарной классификации. Она уже из коробки даёт высокое качество, добиваясь стабильных >0.75 по любым основным метрикам. Визуализации показывают структуру данных и уверенность модели.