

# Анализ привычек студентов и их академической успеваемости

Смирнов Д. В.

Механико-математический факультет  
Московский государственный университет

2025



В проекте будет проводиться анализ датасета:

## Student Habits and Academic Performance Dataset

Будет сделан разведочный анализ данных (EDA). Будут построены и оценены три модели:

- Линейная регрессия
- k ближайших соседей(kNN)
- Случайный лес(Random Forest)

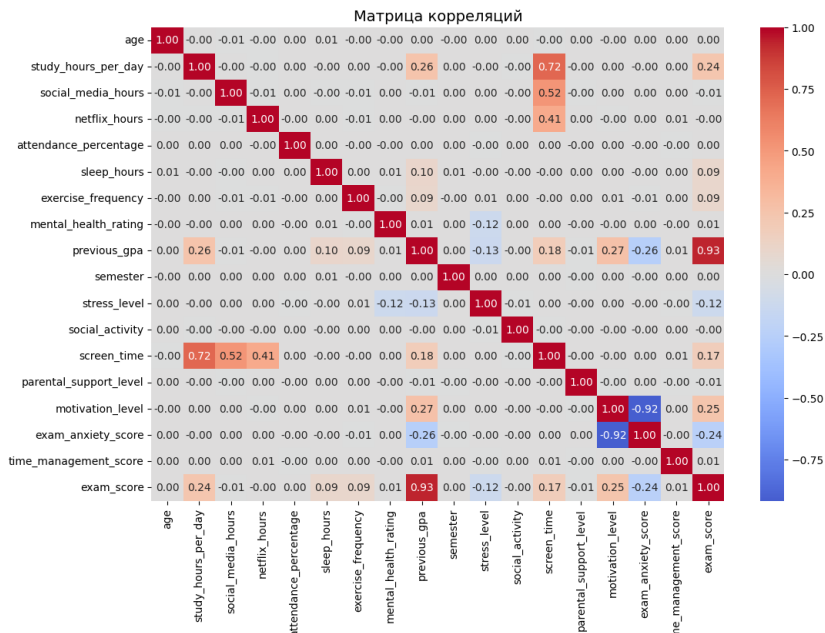
Оцениваться модели будут с помощью 3 метрик:

- $MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$  - среднеквадратичное отклонение
- $MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$  - среднее абсолютное отклонение
- $R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$  - коэффициент детерминации

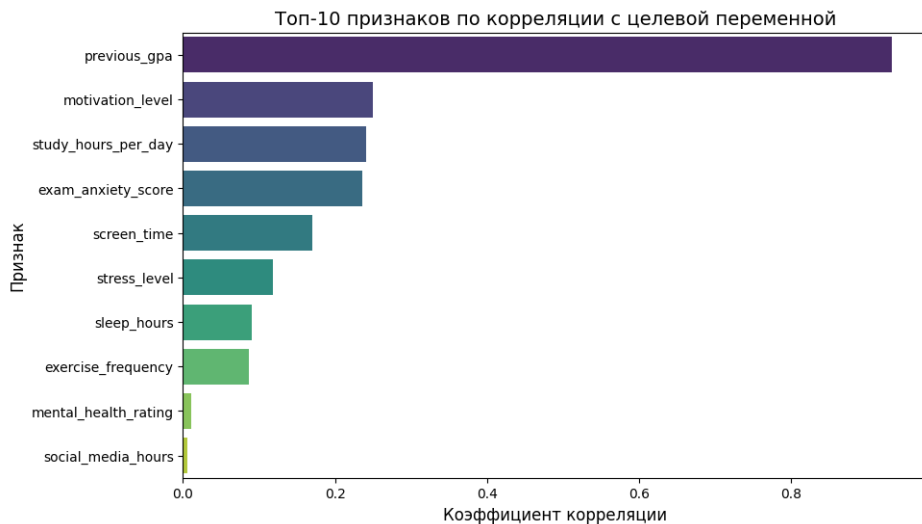
$y_i, \hat{y}_i, \bar{y}$  - истинное значение, предсказанное значение, среднее целевой переменной

$N$  - количество наблюдений в тестовой выборке

Размер датасета  $80000 \times 31$ , причём пропущенных значений нет. Есть целевая переменная **exam\_score** - балл за экзамен. Посмотрим на корреляцию числовых признаков:



А теперь на топ-10 признаков по корреляции с целевой переменной:



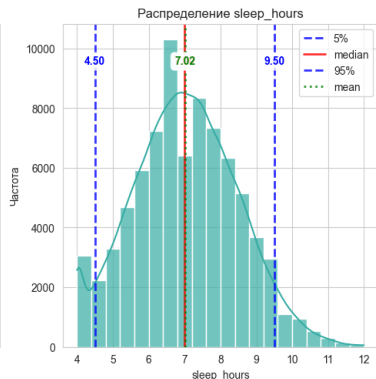
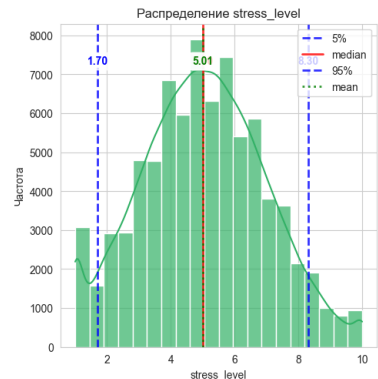
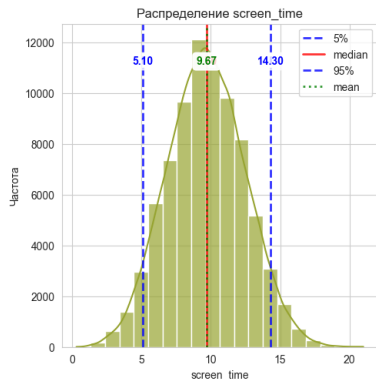
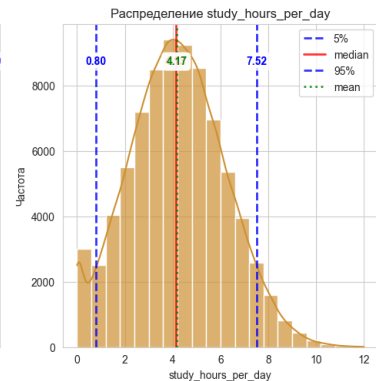
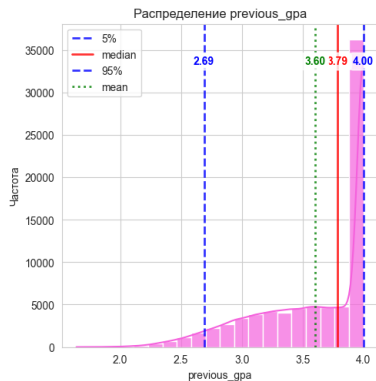
Берём до признака "ментальное здоровье"

Прделаем то же самое с категориальными(кодировали с помощью LabelEncoder):

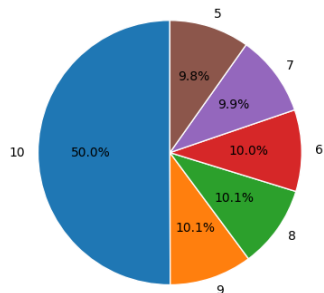


Берём из них: доступ к обучению, обычное место учёбы и риск отчисления

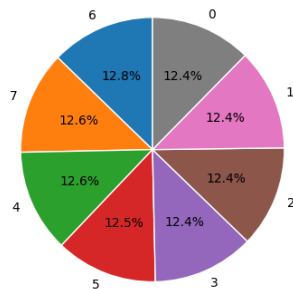
# Распределение числовых признаков



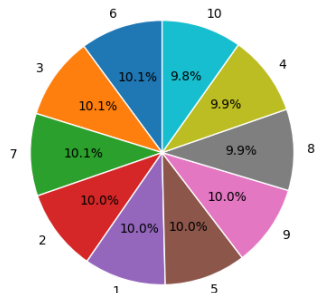
Тревожность на экзамене



Частота физ. упражнений

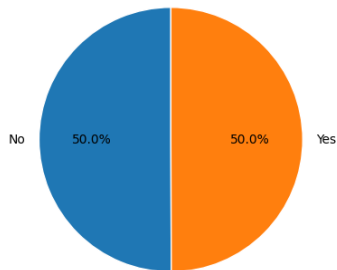


Уровень мотивации

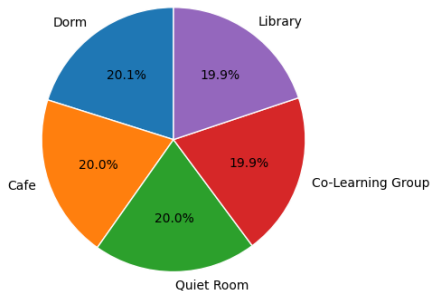


## Распределение категориальных признаков

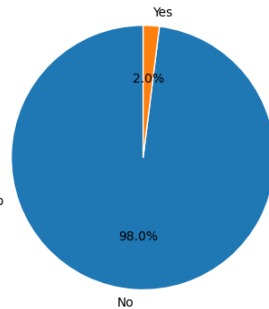
Распределение по доступу к обучения



Распределение по месту обучения



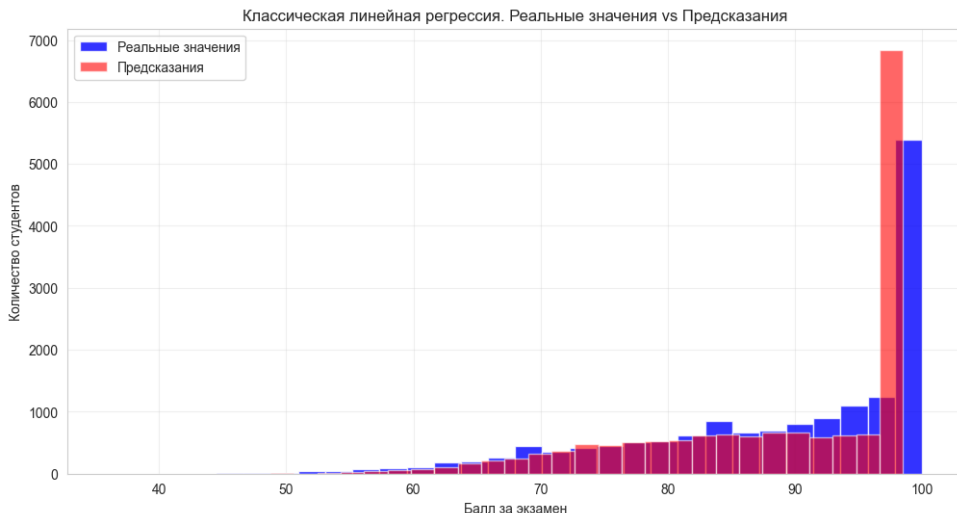
Распределение по риску отчисления



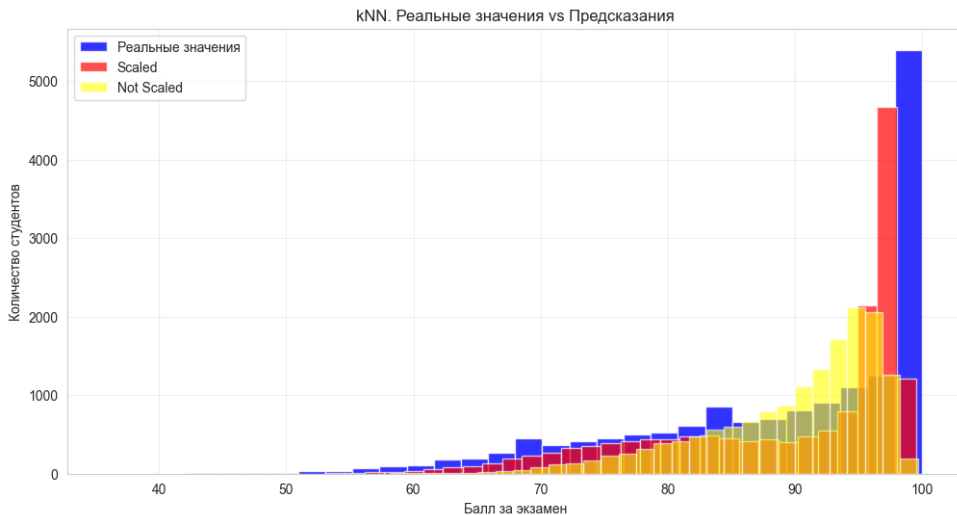


# Линейная регрессия

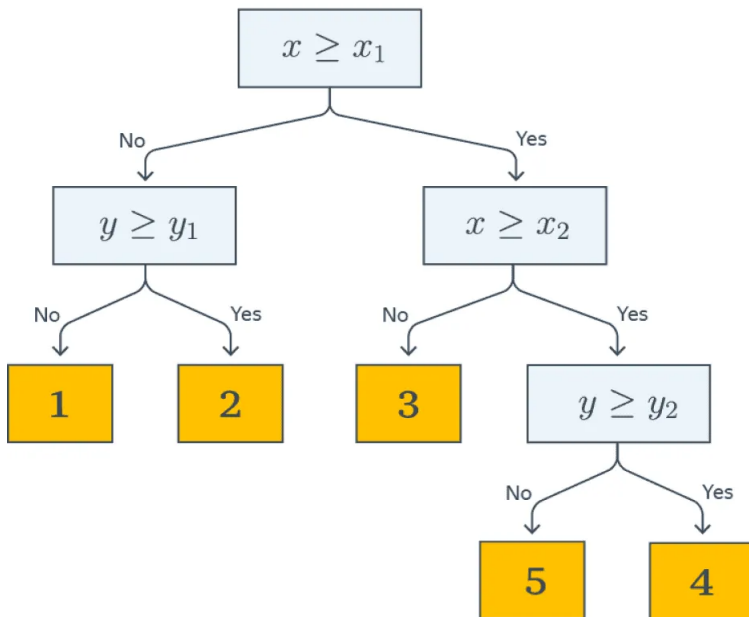
Сравнение качества				
Модификация	Scaled	<i>MAE</i>	<i>MSE</i>	$R^2$
Linear Regression	Yes	3.1919	17.3130	0.8714
	Not	3.1919	17.3130	0.8714
L1-Regularization	Yes	3.1926	17.3138	0.8714
	Not	3.1944	17.3143	0.8714
L2-regularization	Yes	3.1926	17.3138	0.8714
	Not	3.1944	17.3143	0.8714



Сравнение качества				
Модель	Scaled	$MAE$	$MSE$	$R^2$
kNN	Yes	3.6	21.5	0.84
	Not	5.7	52.3	0.61



## Решающее дерево

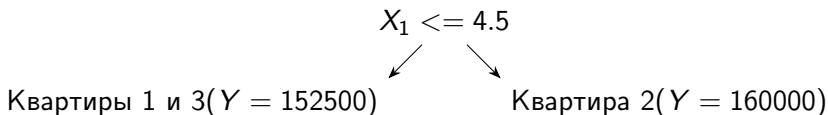


У нас есть 5 наблюдений (квартир) с двумя признаками:

Квартира	X1 (Этаж)	X2 (Площадь, м <sup>2</sup> )	Y (Цена, \$)
1	4	9	150,000
2	5	8	160,000
3	4	10	155,000
4	3	9	145,000
5	4	8	158,000

Выбираем случайно 5 квартир(возможны повторения). Это будет одно дерево

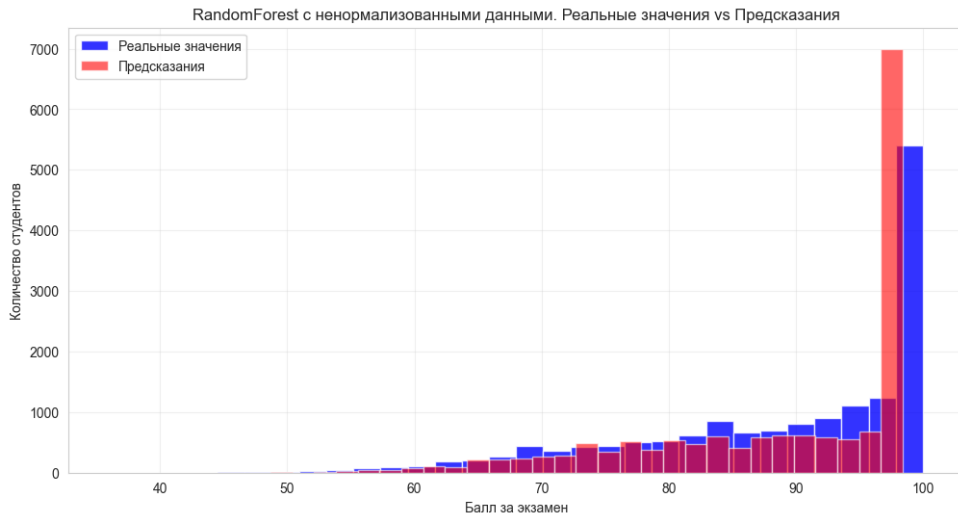
Квартира	X1 (Этаж)	X2 (Площадь, м <sup>2</sup> )	Y (Цена, \$)
1	4	9	150,000
2	5	8	160,000
3	4	10	155,000



Новая квартира(объект из тестовых данных)

Квартира	X1 (Этаж)	X2 (Площадь, м <sup>2</sup> )	Y (Цена, \$)
6	3	12	170,000

Сравнение качества				
Модель	Scaled	MAE	MSE	$R^2$
RandomForest	Yes	3.22	17.1	0.8729
	Not	3.221	17.09	0.873



Сравнение качества моделей			
Модель	$MAE$	$MSE$	$R^2$
LinearRegression	3.1919	17.3130	0.8714
kNN	3.6	21.5	0.84
RandomForest	3.2214	17.1	0.873

