Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ
КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Разработка и оценка моделей машинного обучения

Студент	ИУ5-62Б		**Е.С. Вешторт**
	(группа)	(подпись, дата)	(И.О. Фамилия)
Руководитель НИР			**Ю.Е. Гапанюк**
		(подпись, дата)	(И.О. Фамилия)

2025 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ5
(индекс)
В.И. Терехов
(И.О. Фамилия)
(подпись)
(дата)

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Разработка и оценка моделей машинного обучения

Студент группы ИУ5-62Б
Вешторт Ева

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
ИССЛЕДОВАТЕЛЬСКАЯ

Источник тематики (кафедра, предприятие, НИР) КАФЕДРА

График выполнения НИР:
25% к _____ нед., 50% к _____ нед., 75% к _____ нед., 75% к _____ нед

Техническое задание: *решение задачи машинного обучения на основе материалов* дисциплины. Выбор датасета, первичный анализ, выбор метрик для оценки качества моделей, построение базового решения, оценка качества, подбор гиперпараметров.

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _____ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

Дата выдачи задания « 07 » февраля 2025 г.

Руководитель НИР _____ **Ю.Е. Гапанюк**
(подпись, дата) (И.О. Фамилия)

Студент _____ **Е.С. Вешторт**
(подпись, дата) (И.О. Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ **4**
1. Обзор данных и предобработка **5**
2. Разведочный анализ данных (EDA) **6**
3. Моделирование **9**
4. Веб-приложение ..... Ошибка! Закладка не определена.
ЗАКЛЮЧЕНИЕ **12**
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ **13**

ВВЕДЕНИЕ

В данном исследовании анализируется набор данных "Student Habits vs Academic Performance", содержащий данные о привычках и качествах студентов и финальном балле на экзамене.

Целью работы является:

1. Понимание факторов, влияющих на результат на экзамене.
2. Построение моделей машинного обучения для решения двух задач:
   - Задача регрессии: предсказание точного балла студента на экзамене.
   - Задача классификации: предсказание категории студента (лучшие, средние, отстающие, деление по перцентилям).

Набор данных включает информацию о возрасте студента, количестве часов, потраченных на занятия, частоту посещения занятий, качество интернета, уровень образования родителей и другие.

1. Обзор данных и предобработка

Загрузка и начальный обзор данных:

- Загрузка датасета с использованием библиотеки pandas.
- Отображение первых строк данных (.head()).
- Проверка типов данных и наличия пропущенных значений (.info()).

	student_id	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
0	S1000	23	Female	0.0	1.2	1.1	No	85.0	8.0	Fair	6	Master	Average	8	Yes	56.2
1	S1001	20	Female	6.9	2.8	2.3	No	97.3	4.6	Good	6	High School	Average	8	No	100.0
2	S1002	21	Male	1.4	3.1	1.3	No	94.8	8.0	Poor	1	High School	Poor	1	No	34.3
3	S1003	23	Female	1.0	3.9	1.0	No	71.0	9.2	Poor	4	Master	Good	1	Yes	26.8
4	S1004	19	Female	5.0	4.4	0.5	No	90.9	4.9	Fair	3	Master	Good	1	No	66.4

Заполнение пропусков:

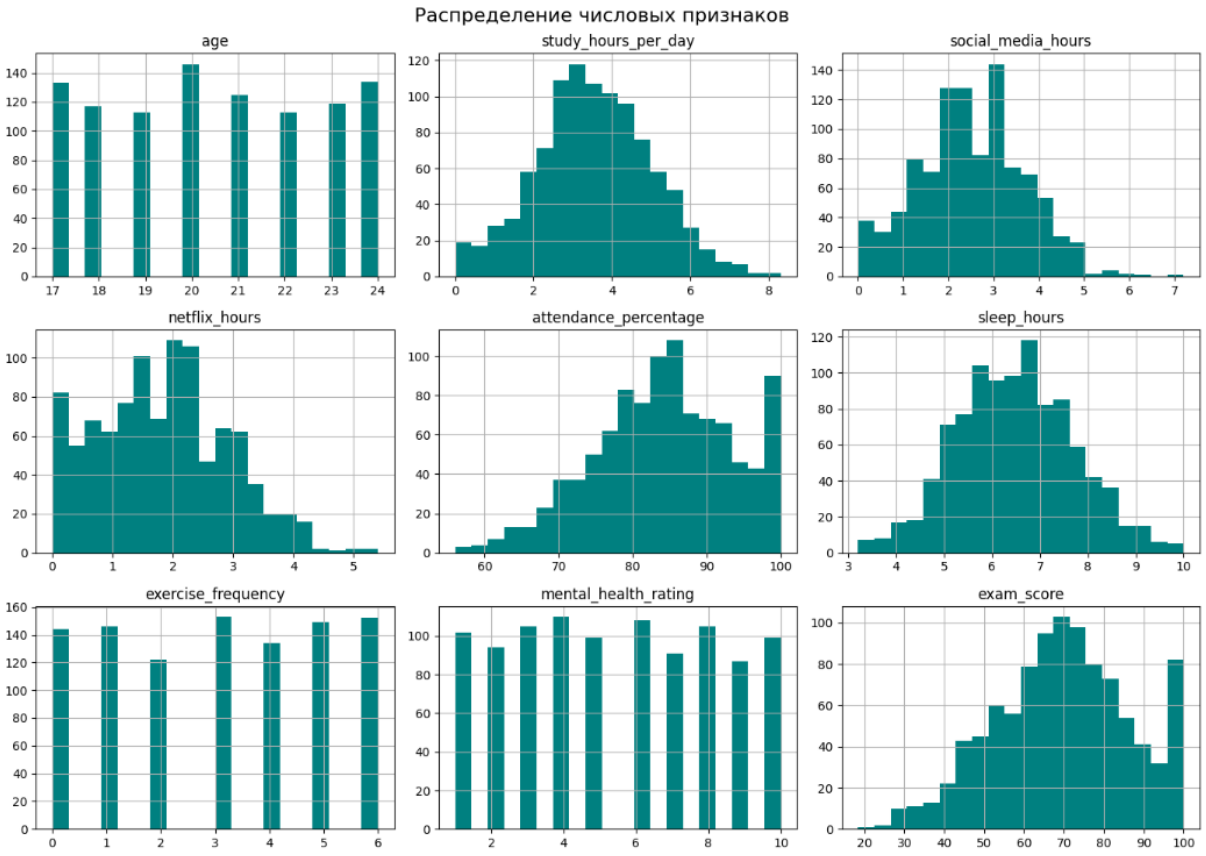
В датасете все данные приведены в нужных типах и пропуски есть только в столбце с информацией об образовании родителей. Заполним пропуски строкой “Unknown”

Обработка категориальных признаков:

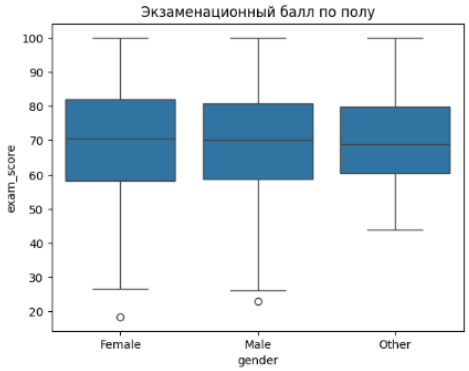
- Применение One-Hot Encoding (OHE) для категориальных признаков.

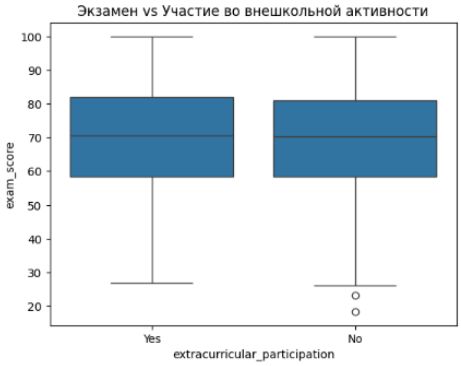
2. Разведочный анализ данных (EDA)

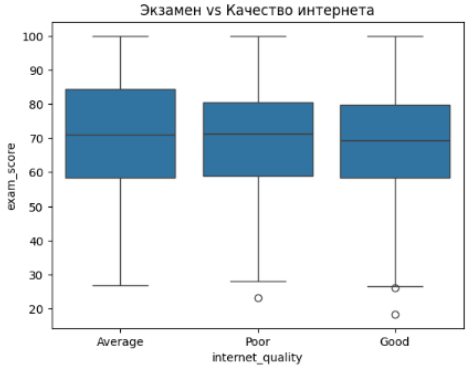
Распределения числовых признаков:

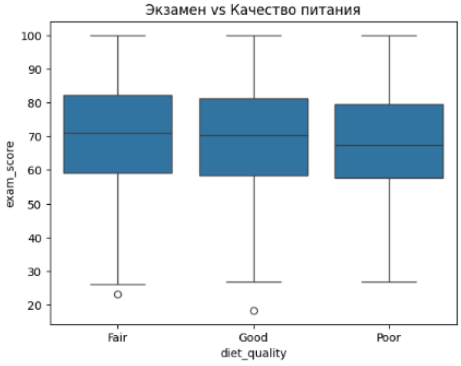


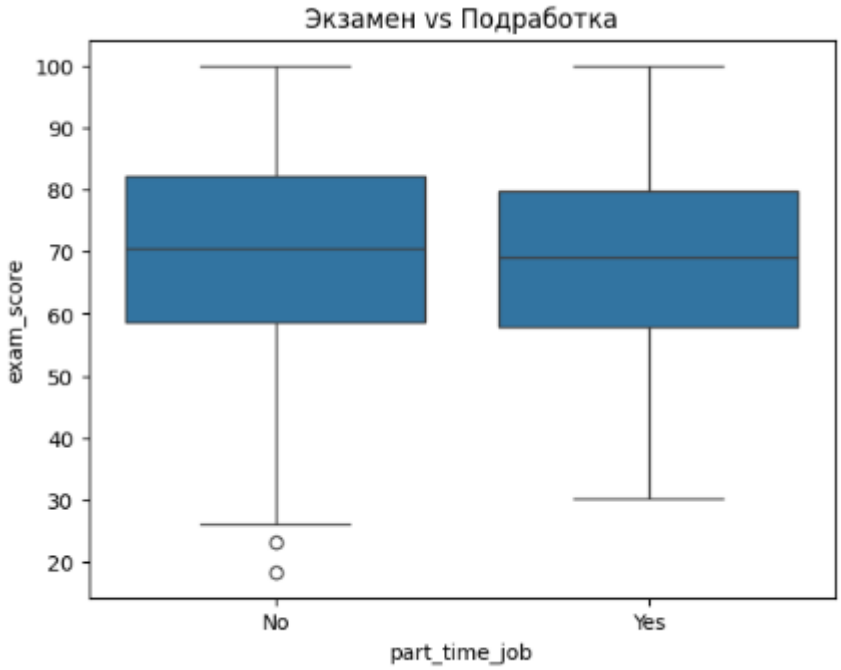
Влияние категориальных признаков на целевую переменную:



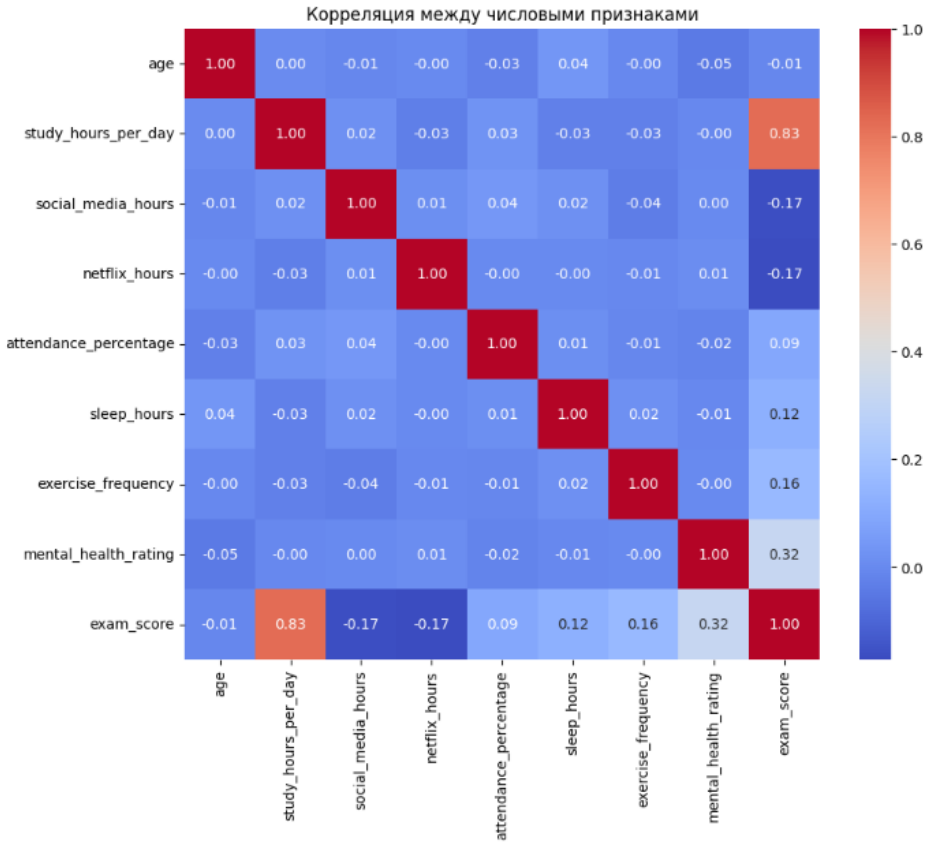








Корреляция между числовыми признаками:



Выводы:

- Наиболее важные факторы, положительно влияющие на результат экзамена это количество часов занятий в день, уровень

ментального здоровья, частота упражнений и количество часов сна.

- Наиболее важные факторы, отрицательно влияющие на оценку на экзамене, это количество часов просмотра сериалов и социальных сетей.
- Категориальные признаки, такие как пол или уровень образования родителей, слабо влияют на результат.

3. Моделирование

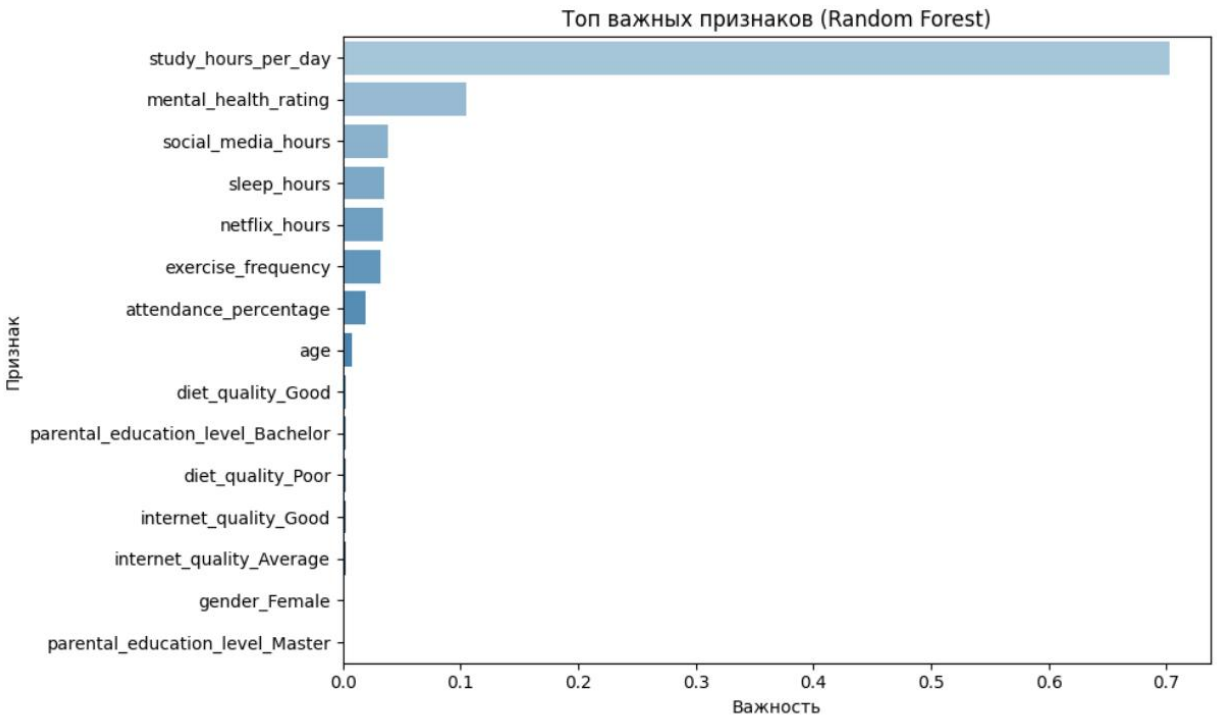
Для обеих задач (регрессии и классификации) данные были разделены на обучающую и тестовую выборки (80/20). Использовались различные модели машинного обучения, а также проведена оптимизация гиперпараметров с помощью GridSearchCV и кросс-валидации.

4.1. Задача регрессии: Предсказание точного балла на экзамене

- **Используемые модели:** Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor.
- **Метрики оценки:** R2 (коэффициент детерминации), MAE (средняя абсолютная ошибка), RMSE (среднеквадратичная ошибка).
- **Результаты:**
  - Наилучшие результаты показали ансамблевые модели: Random Forest Regressor и Gradient Boosting Regressor.

	Model	R2	MAE	RMSE
2	Random Forest	0.980915	1.864820	2.331980
3	Gradient Boosting	0.913800	3.997694	4.955988
0	Linear Regression	0.901845	4.200019	5.288477
1	Decision Tree	0.829565	5.534503	6.968745

  - После подбора гиперпараметров удалось достичь R2 0.98, MAE и RMSE в районе 2 баллов. Это означает, что лучшая модель способна объяснить около 98 дисперсии признаков.
- **Важность признаков:**
  - Важность признаков, полученная из RandomForest, подтвердила выводы EDA: наиболее значимые признаки это количество часов учебы и ментальное здоровье.

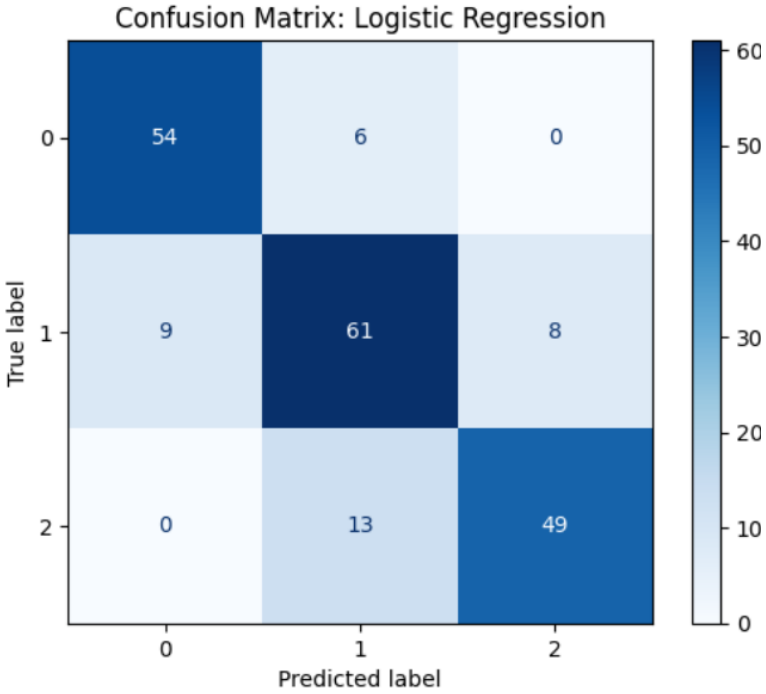


4.2. Задача классификации: Предсказание категории студента (Отстающий/Средний/Успешный)

- **Создание целевой переменной:** Количество арендованных велосипедов было разделено на три категории по квантилям (отстающий, средний, успешный).
- **Используемые модели:** Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, KNeighbours.
- **Метрики оценки:** F1-score (weighted), Accuracy, Precision, Recall, Confusion Matrix.
- **Результаты:**
  - Лучше всего себя показала регрессия.

	Model	Accuracy	F1_score	Precision	Recall
1	Logistic Regression	0.820	0.819848	0.821009	0.820
4	Gradient Boosting	0.755	0.751163	0.751739	0.755
2	Decision Tree	0.725	0.724855	0.726901	0.725
3	Random Forest	0.720	0.714568	0.714946	0.720
0	KNearest	0.705	0.706366	0.709439	0.705

- После тюнинга удалось достичь F1-score (weighted) около 0.82, Accuracy около 0.82.
- Confusion matrix показала хорошее предсказание основной диагонали, но наблюдались некоторые ошибки между смежными классами.



ЗАКЛЮЧЕНИЕ

В ходе выполнения научно-исследовательской работы был проведен анализ данных о результатах студентов на экзамене и построены модели машинного обучения для задач регрессии и классификации.

- Проведен разведочный анализ данных, который выявил ключевые факторы, влияющие на спрос, такие как время учебы в день, уровень ментального здоровья, частоту упражнений.
- Для задачи регрессии ансамблевые модели (Random Forest Regressor, Gradient Boosting Regressor) показали высокую точность предсказания.
- Для задачи классификации (разделение студентов на категории) наиболее высокую точность показала модель регрессии.

Результаты исследования могут быть использованы для мотивации студентов и улучшения их учебных показателей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Student Habits vs Academic Performance [Электронный ресурс] // github.com. URL: https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance/data (дата обращения: 22.05.2025);
2. Документация Streamlit [Электронный ресурс] // streamlit.io URL: https://streamlit.io/ (дата обращения: 01.05.2025);
3. «Python Data Science Handbook» Джейк Вандер-Плас [Электронный ресурс] // jakevdp.github.io. URL: https://jakevdp.github.io/PythonDataScienceHandbook/ (дата обращения: 02.05.2025);
4. Документация по Python [Электронный ресурс] // Python. URL: https://docs.python.org/3/index.html/ (дата обращения: 01.05.2025);
5. Методические указания НИРС по дисциплине «Технологии машинного обучения».