



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 веб-приложения для построения
моделей машинного обучения*

| | | | |
|------------------|---------------|-----------------|---------------------|
| Студент | <hr/> ИУ5-65Б | <hr/> | <hr/> Ходырев Р. В. |
| | (группа) | (подпись, дата) | (И.О. Фамилия) |
| Руководитель НИР | | <hr/> | <hr/> 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)

ЗАДАНИЕ

на выполнение научно-исследовательской работы

по теме Разработка веб-приложения для построения моделей методов
машинного обучения

Студент группы ИУ5-65Б

Ходырев Роман Владиславович

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:

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

(подпись, дата)

Ю.Е. Гапанюк

(И.О. Фамилия)

Студент

(подпись, дата)

Ходырев Р. В.

(И.О. Фамилия)

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

Оглавление

Введение4

Постановка задачи4

Подбор и подготовка данных5

Исследовательский анализ данных (EDA)6

Построение и сравнение моделей11

История.....14

Заключение.....15

Список использованных источников15

Введение

Машинное обучение на сегодняшний день является одним из наиболее активно развивающихся направлений в области анализа данных. Оно широко применяется для решения различных задач, включая классификацию, регрессию, кластеризацию и многое другое. В данной работе разрабатывается веб-приложение для анализа и/или построения моделей машинного обучения на основе представленного числового датасета.

Целью настоящего исследования является построение, обучение и сравнение нескольких моделей машинного обучения, включая ансамблевые методы, а также проведение полного цикла подготовки и анализа данных. В процессе работы производится оценка качества моделей по различным метрикам, формирование выводов о качестве решений, а также демонстрация полученного результата в виде веб-приложения.

Постановка задачи

Имеется несколько открытых наборов данных, содержащих числовые величины. На основе этих данных веб-приложению необходимо проанализировать датасет, построив 3 вида диаграмм – парную для каждого параметра, скрипичные диаграммы и корреляционную матрицу. Также есть возможность построить модели машинного обучения (8 моделей, из которых 3 ансамблевых) для предусмотренных для этого датасетов.

Для решения задачи требуется:

- Выполнить разведочный анализ данных;
- Обработать пропуски и закодировать категориальные переменные;
- Провести масштабирование признаков;
- Сформировать обучающую и тестовую выборки;
- Построить не менее пяти моделей (включая две ансамблевые);
- Оценить их качество по нескольким метрикам (R^2 и MAE);
- Сравнить результаты и обосновать выбор финальной модели.

Подбор и подготовка данных

В качестве примера работы веб-приложения выберем датасет Predict Student Performance, представленный в виде CSV-файла students.csv. Набор включает такие признаки, как:

- Учебные часы - Среднее количество ежедневных часов, потраченных на учебу.
- Часы сна - Среднее количество ежедневных часов, потраченных на сон.
- Социально-экономический балл - Нормализованный балл (0-1), указывающий на социально-экономическое положение учащегося.
- Посещаемость (%) - Процент занятий, которые посещал студент.
- Оценки (ЦЕЛЕВАЯ) - Итоговый балл успеваемости студента, полученный на основе комбинации учебных часов, часов сна, социально-экономического балла и посещаемости.

Для подготовки данных были выполнены следующие шаги:

1. **Исключение пропусков:**
2. **Удаление нерелевантных признаков:** при наличии нечисловых признаков они также удаляются

Главная страница

На главной странице приложения есть 3 доступных кнопки – «Анализ датасета», «Построение моделей» и «История».

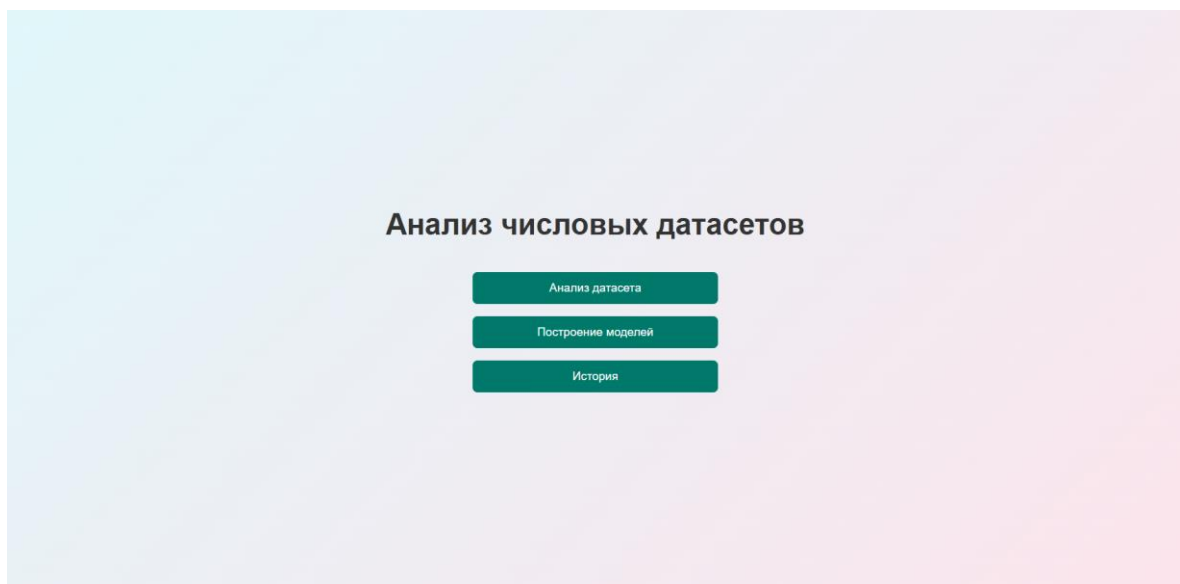


Рисунок 1 – Главная страница веб-приложения

Рассмотрим подробнее каждый пункт.

Исследовательский анализ данных (EDA)

Разведочный анализ данных позволил выявить особенности распределения признаков и их связь с целевой переменной. Это важный этап, позволяющий сформировать гипотезы и принять решения по обработке и отбору признаков.

Для начала анализа датасета на главной странице нажимаем кнопку «Анализ датасета».

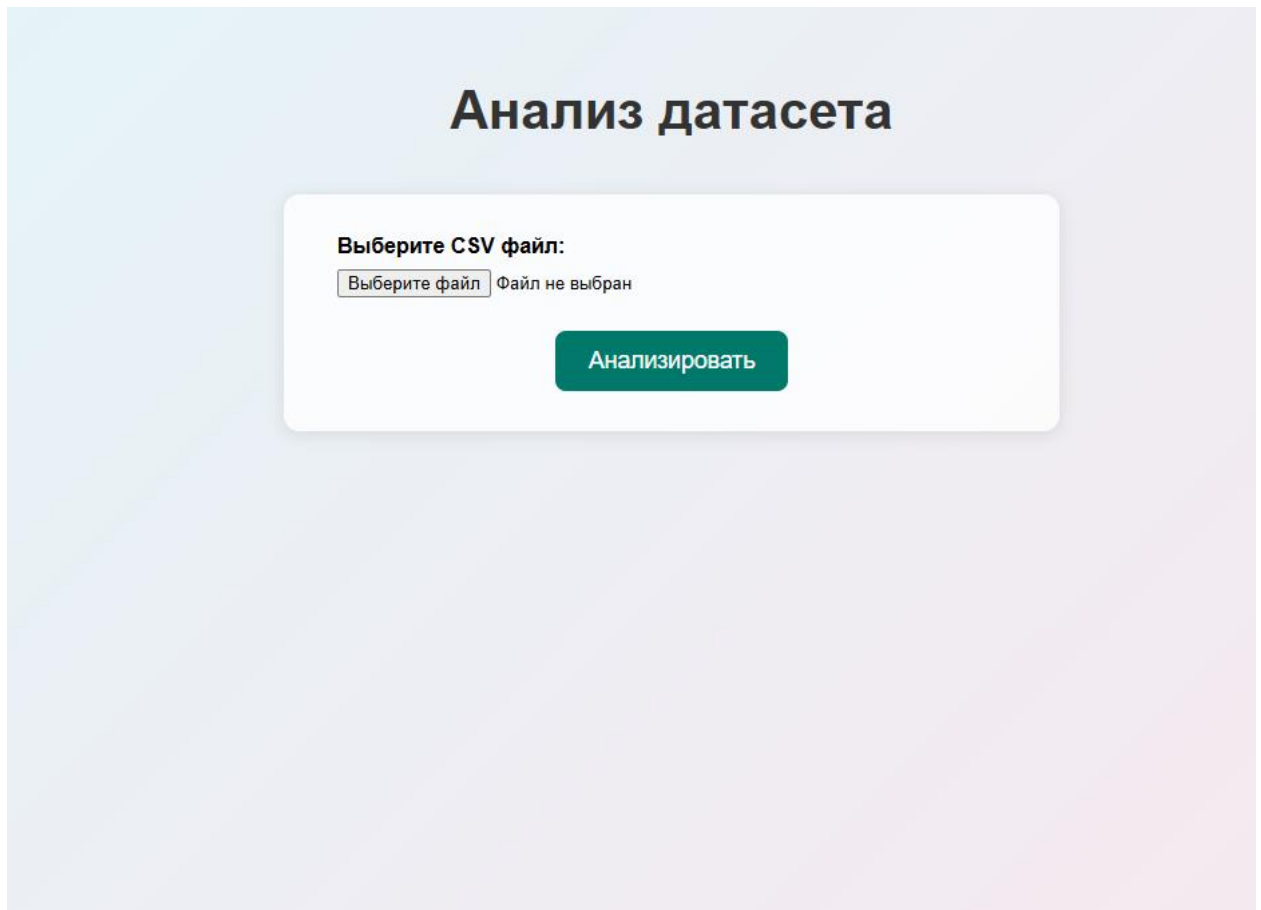


Рисунок 2 – Анализ датасета

Необходимо нажать на кнопку «Выберите файл» и загрузить .csv файл со своего устройства. Датасет должен содержать только числовые признаки. После выбора датасета Predict Student Performance нажимаем кнопку «Анализировать».

Посмотрим на результаты анализа датасета.

Анализ датасета

Типы данных

| Колонка | Тип |
|---------------------|---------|
| Socioeconomic Score | float64 |
| Study Hours | float64 |
| Sleep Hours | float64 |
| Attendance (%) | float64 |
| Grades | float64 |

Рисунок 3 – Результаты анализа датасета. Типы данных

Сначала приложение анализирует все колонки датасета и тип данных. В нашем случае все поля имеют тип float64 или число с плавающей запятой.

Первые 10 строк

| Socioeconomic Score | Study Hours | Sleep Hours | Attendance (%) | Grades |
|---------------------|-------------|-------------|----------------|--------|
| 0.95822 | 3.4 | 8.2 | 53.0 | 47.0 |
| 0.85566 | 3.2 | 5.9 | 55.0 | 35.0 |
| 0.68025 | 3.2 | 9.3 | 41.0 | 32.0 |
| 0.25936 | 3.2 | 8.2 | 47.0 | 34.0 |
| 0.60447 | 3.8 | 10.0 | 75.0 | 33.0 |
| 0.9832 | 3.4 | 9.0 | 47.0 | 51.0 |
| 0.56648 | 7.9 | 8.1 | 63.0 | 54.0 |
| 0.93487 | 1.4 | 8.0 | 47.0 | 34.0 |
| 0.4666 | 5.4 | 8.8 | 67.0 | 39.0 |
| 0.6213 | 1.4 | 9.6 | 42.0 | 34.0 |

Рисунок 4 – Результаты анализа датасета. Первые 10 строк

Далее для удобства пользователя выводятся первые 10 строк загруженного датасета. Это позволяет визуально оценить датасет, понять важность полей, их примерные диапазоны.

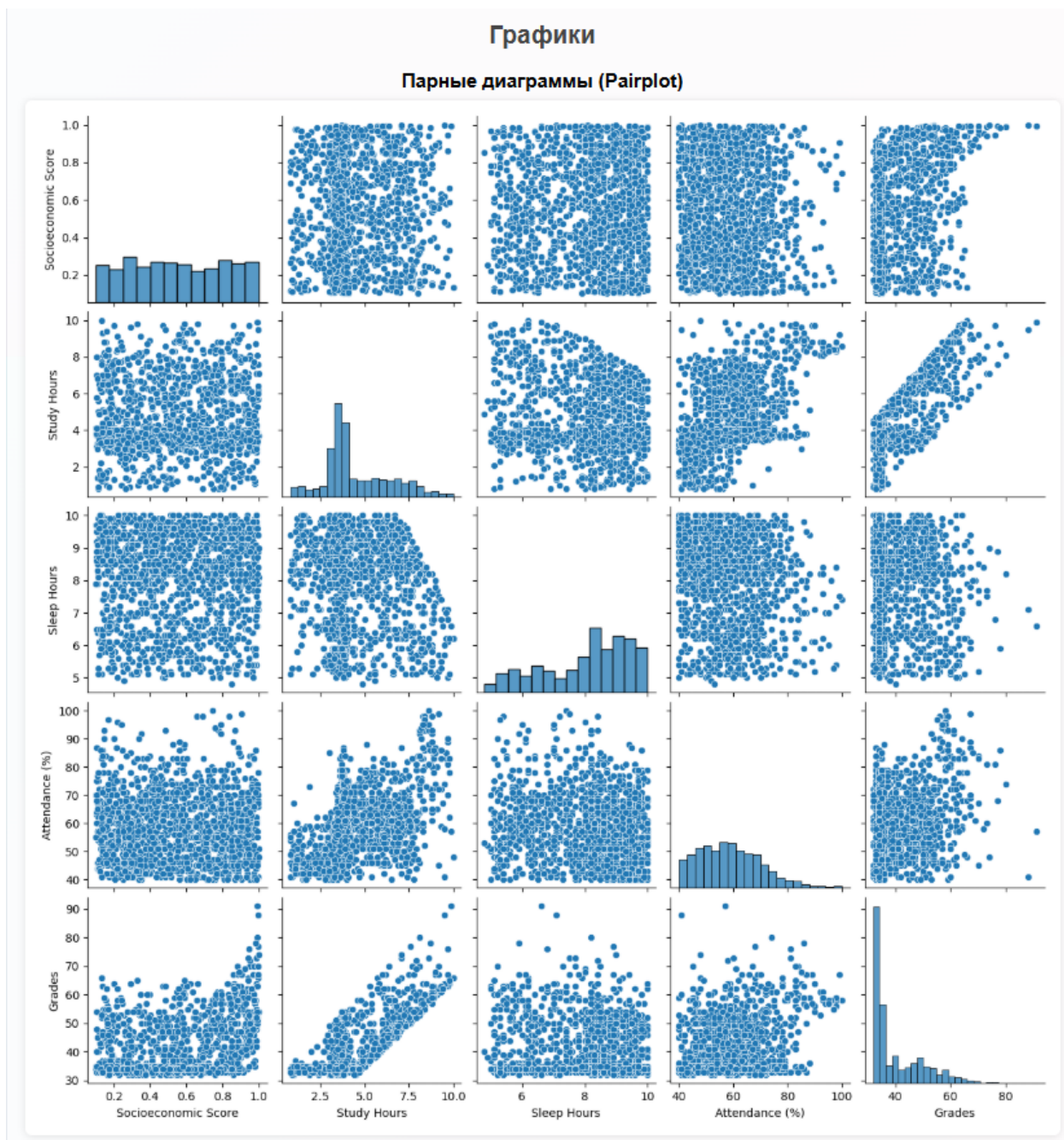


Рисунок 5 – Результаты анализа датасета. Парные диаграммы

На парных диаграммах можно оценить связь всех переменных между собой, оценить какие данные более важные для датасета, а какие, наоборот, можно не учитывать.

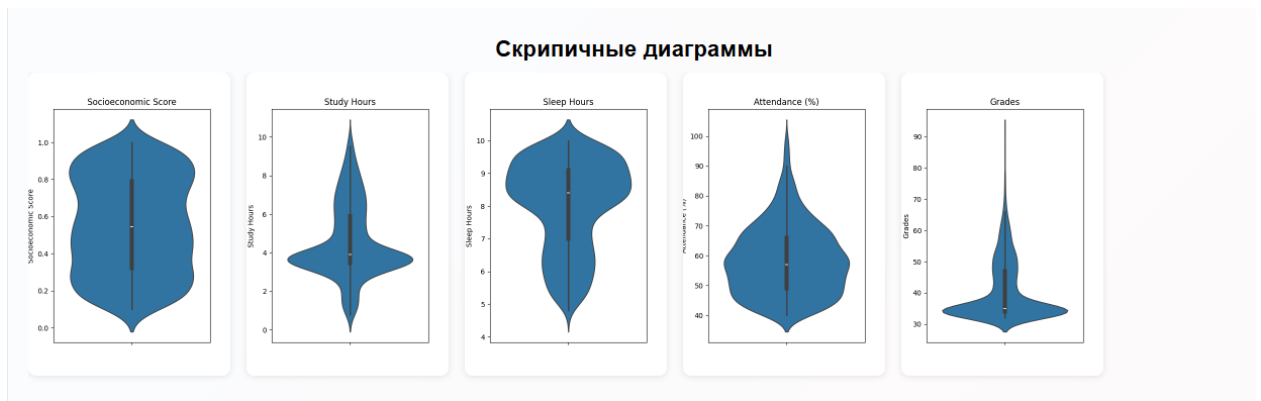


Рисунок 6 – Результаты анализа датасета. Скрипичные диаграммы

Далее представлены скрипичные диаграммы, показывающие как распределены данные в одном поле. Понять средние значения, а также количество максимальных и минимальных значений для одной колонки

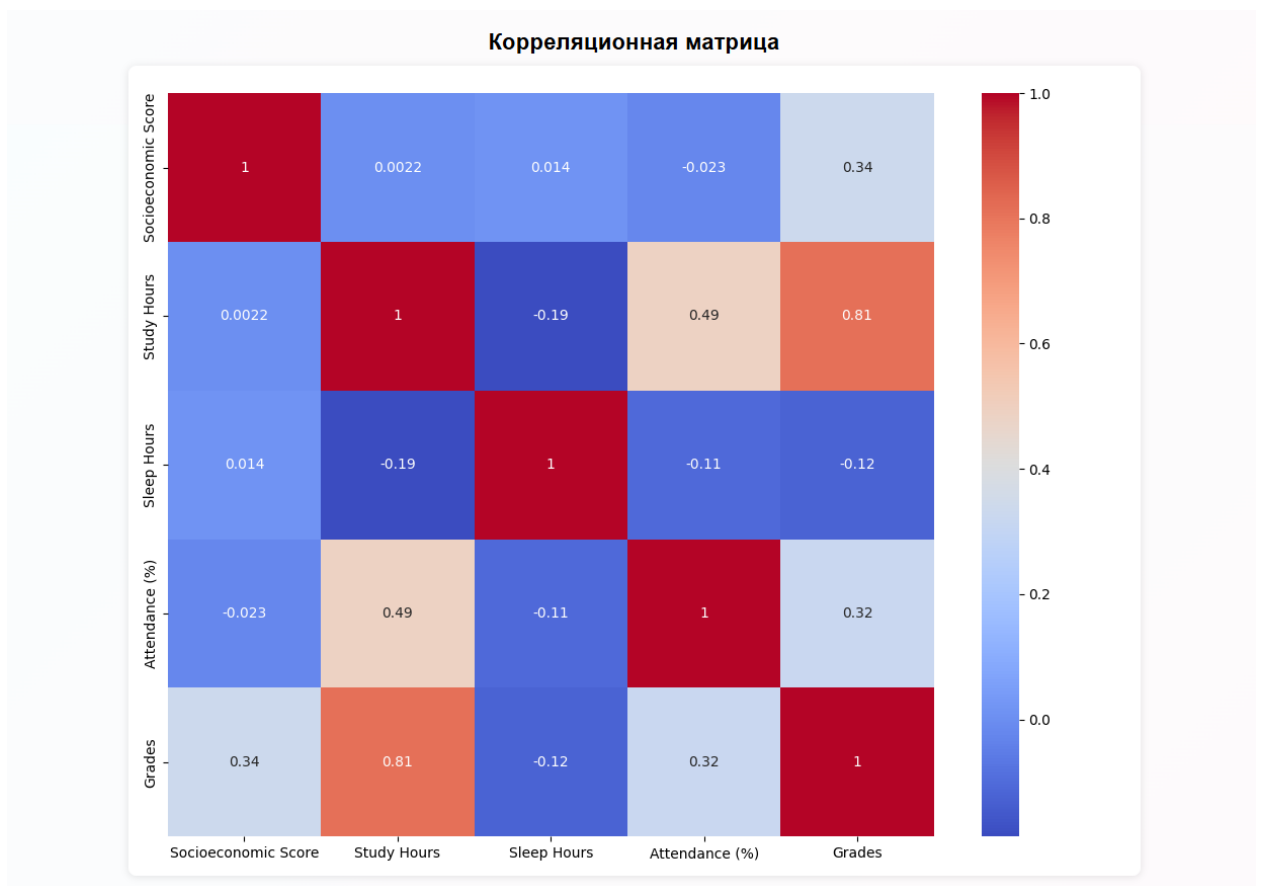


Рисунок 7 – Результаты анализа датасета. Корреляционная матрица

Матрица корреляций показывает взаимосвязи между признаками в датасете Оценок студентов:

Сильные корреляции:

- Study Hours и Grades (0.81) – оценка студента сильно зависит от учебных часов.
- Attendance и Study Hours (0.49) — посещаемость зависит от учебных часов.

Умеренные корреляции:

- Socioeconomic Score и Grades (0.34) — социально-экономический показатель студента влияет на его оценку.
- Attendance и Grades (0.32) — в некоторых случаях посещаемость влияет на оценку
- Study Hours и Sleep Hours (-0.19) — чем больше у студента учебных часов, тем меньше спит студент, и наоборот.

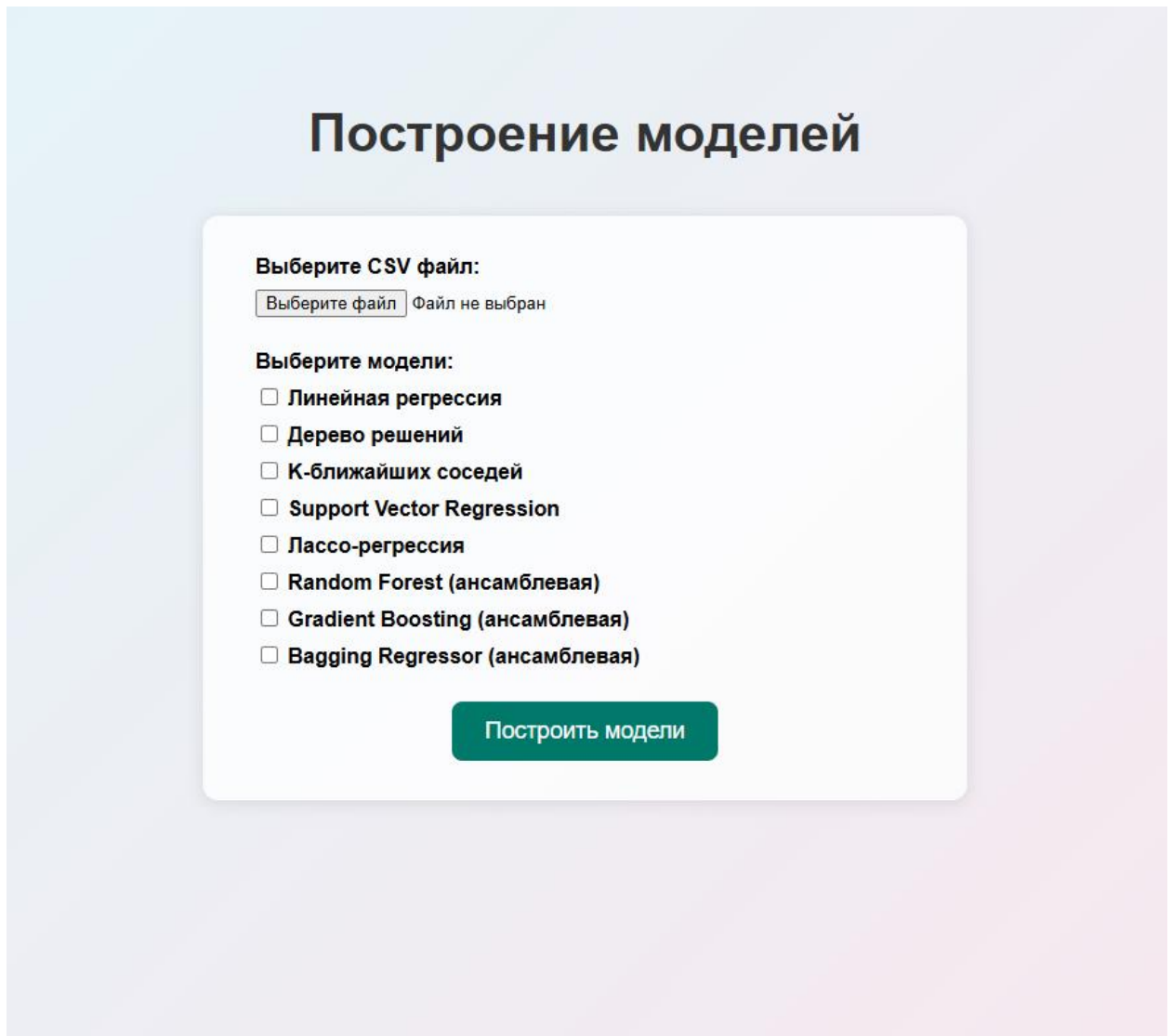
Слабые корреляции:

- Большинство остальных пар имеют корреляцию близкую к нулю, что указывает на их независимость.

При нажатии на кнопку «Назад» происходит переход на главный экран

Построение и сравнение моделей

При нажатии на главной странице на кнопку «Построение моделей» откроется следующая страница:



Построение моделей

Выберите CSV файл:

Выберите файл Файл не выбран

Выберите модели:

- ☐ Линейная регрессия
- ☐ Дерево решений
- ☐ K-ближайших соседей
- ☐ Support Vector Regression
- ☐ Лассо-регрессия
- ☐ Random Forest (ансамблевая)
- ☐ Gradient Boosting (ансамблевая)
- ☐ Bagging Regressor (ансамблевая)

Построить модели

Рисунок 8 – Построение моделей

На этой странице пользователю предлагается выбрать датасет также в формате .csv файла, состоящего из числовых признаков. Для примера также выберем датасет Predict Student Performance.

Пользователю предлагается построение и обучение следующих модели:

1. Линейная регрессия
2. Дерево решений
3. KNN
4. Support Vector Regression
5. Лассо-регрессия

6. Случайный лес
7. Градиентный бустинг
8. Bagging Regressor

Все модели обучаются на одной и той же выборке (80% обучающая, 20% тестовая).

Выберем все модели и построим их.

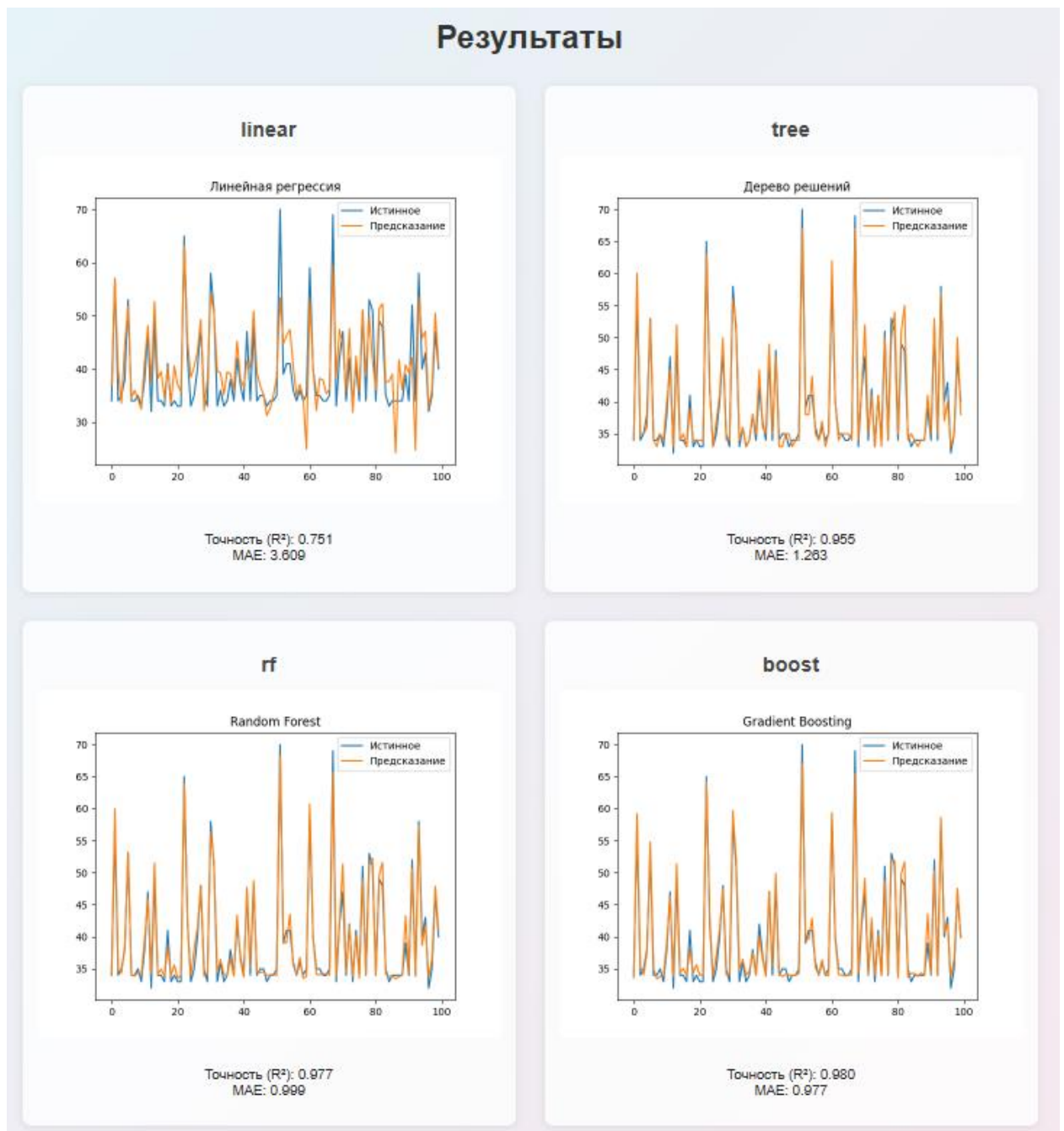


Рисунок 9 – Построение моделей. Результат. Часть 1

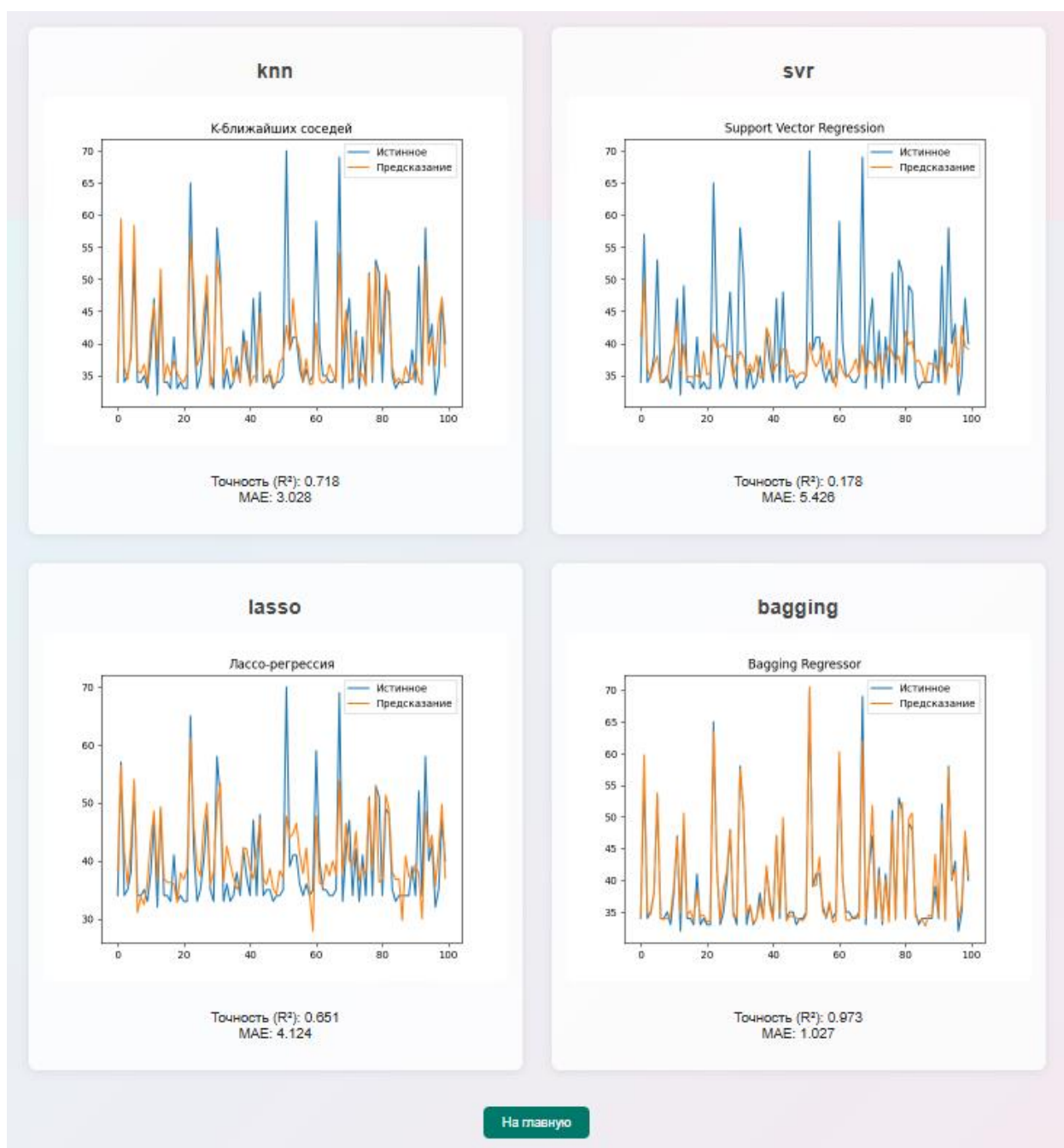


Рисунок 10 – Построение моделей. Результат. Часть 2

Как видно из результатов построения моделей, лучшую точность показывают ансамблевые модели – Случайный лес, Градиентный бустинг и Bagging Regressor. Это подтверждают метрики R^2 и MAE.

По нажатию кнопки «На главную» происходит переход на главную страницу.

История

Последняя страница – История. На неё можно попасть, если нажать на кнопку «История» на главной странице.

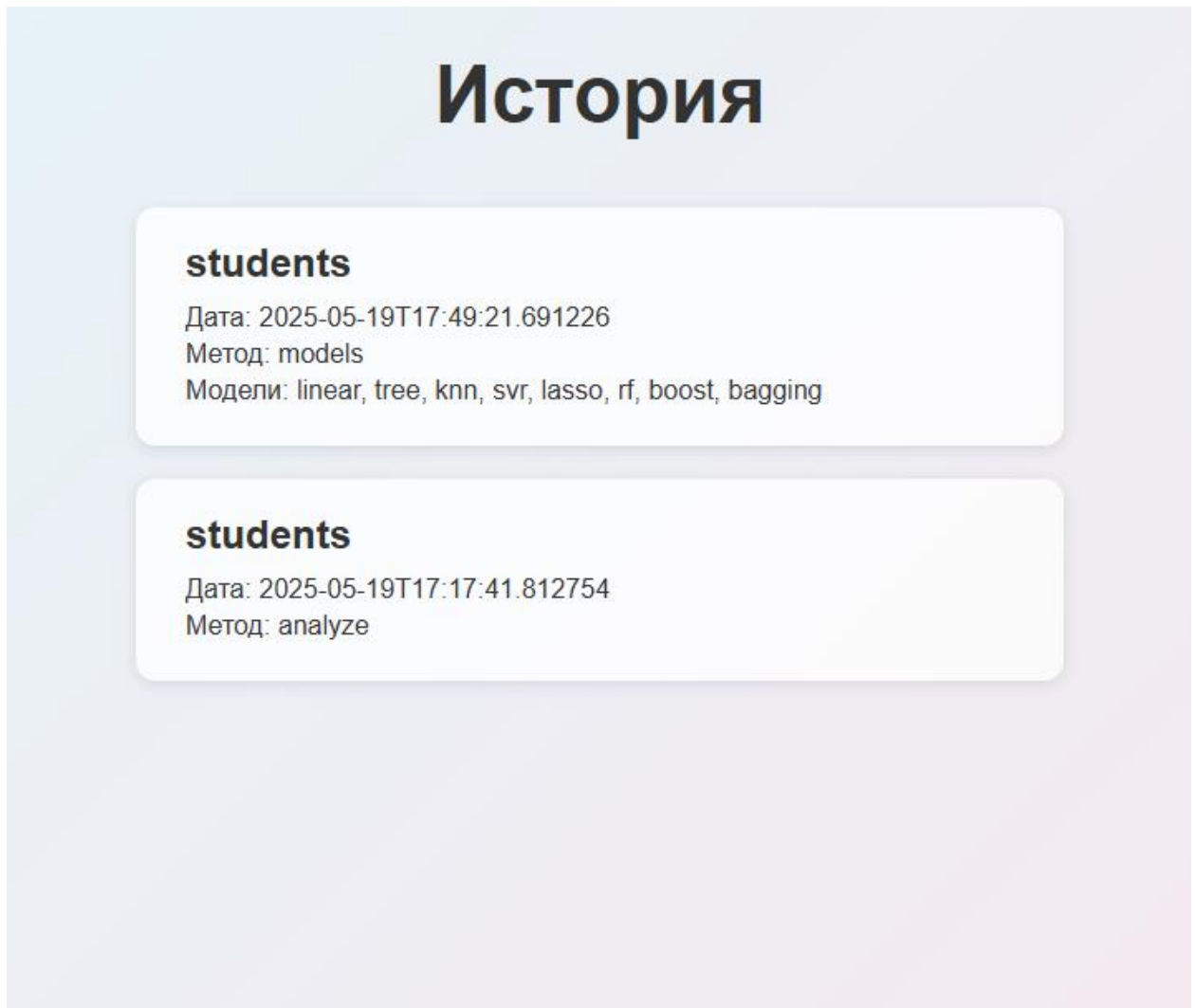


Рисунок 11 – История

На данной странице представлена история действий с веб-приложением. Преимущество такого подхода – отсутствует необходимость повторного анализа или построение моделей. Приложение сохраняет все числовые данные в формате JSON в папке history, а скрины – в папке media. Таким образом, при открытии любой карточки на данной странице приложение просто возьмет закешированные данные и вернёт пользователю.

Заключение

Результаты настоящей работы показали, что при системном подходе к решению задачи классификации можно существенно повысить точность моделей за счёт следующих факторов:

- Тщательная обработка и расширение признаков;
- Корректное кодирование и масштабирование данных;
- Использование продвинутых моделей и подбор гиперпараметров;
- Анализ метрик, позволяющих делать взвешенные выводы.

Наилучшими моделями были признаны ансамблевые модели, но с небольшой точностью всё же выигрывает **Gradient Boosting**, которая показала точность 0.98 при средней абсолютной ошибке меньше 1%. Это делает ее наиболее подходящей для решения поставленной задачи.

Полученные модели и подготовленный код могут быть легко адаптированы для других задач классификации, что демонстрирует универсальность применённого подхода. Кроме того, была реализована сохранённая модель и масштабировщик, которые могут использоваться в продуктивной среде, а также возможно их внедрение в веб-интерфейс с использованием django.

Список использованных источников

1. Kaggle: Predict Student Performance
(<https://www.kaggle.com/datasets/stealthtechnologies/predict-student-performance-dataset>)
2. Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* — O'Reilly, 2019.
3. Документация Scikit-learn — <https://scikit-learn.org/>
4. Материалы курса "Машинное обучение", OpenAI, Stepik, Coursera
5. Python Software Foundation — <https://www.python.org/>
6. Визуализация и EDA: <https://seaborn.pydata.org/>, <https://matplotlib.org/>