|  |  |
| --- | --- |
|  | **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)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_\_РАДИОТЕХНИЧЕСКИЙ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_\_Разработка и оценка моделей\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_РТ5-61Б\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_ А.Е. Фруктин\_\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_****Ю.Е. Гапанюк\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 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)

«\_07\_» \_\_\_\_февраля\_\_\_\_\_\_\_ 2024 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_Разработка и оценка моделей машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_РТ5-61Б\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Фруктин Артём Евгеньевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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., 100% 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\_\_\_\_\_\_\_ 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_Ю.Е. Гапанюк\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_\_\_\_\_А.Е.Фруктин\_\_**\_\_**

(Подпись, дата) (И.О.Фамилия)

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

СОДЕРЖАНИЕ

[Введение 3](#_Toc166636917)

[Основная часть 5](#_Toc166636918)

[Заключение 14](#_Toc166636919)

### Введение

В данной работе мы сосредоточимся на анализе набора данных по успеваемости студентов, предоставленного Kaggle. Этот набор данных включает информацию о предыдущих баллах за тестирование, часов на сон и учебу, вне групповых активностях. Целью работы является разработка и оценка моделей машинного обучения для прогнозирования производительности студентов.

Модели машинного обучения могут предложить ценный инструмент для анализа сложных взаимосвязей между различными признаками и их влиянием на производителньость. Они могут помочь выявить, какие факторы являются наиболее важными для эффективности обучающихся.

В ходе работы будет детальный анализ данных, включая предварительный исследовательский анализ, выбор и кодирование признаков, корреляционный анализ и подбор гиперпараметров для моделей машинного обучения. Я также сравню различные модели, чтобы определить, какая из них наиболее эффективна в прогнозировании индекса результативности.

Результаты работы могут служить полезным руководством для студентов, стремящихся улучшить свою эффективность, а также для исследователей и специалистов в области образования, интересующихся вопросами, влияющими на процесс обучения.

### Основная часть

**Постановка Задачи**

В данном проекте мы будем использовать набор данных о показателях студентов, доступный по ссылке: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression

Цель данной задачи регрессии — предсказать значение производительности (`Performance Index`) на основе предоставленных признаков.

Датасет включает один файл `Student\_Performance.csv` с такими колонками:

- Hours Studied: часы, потраченные на учёбу.

- Previous Scores: Предыдущие результаты за тесты.

- Extracurricular Activities: вне групповые активности.

- Sleep Hours: часы, потраченные на сон.

- Sample Question Papers Practiced: количество образцов вопросов, которые студент отработал.

**Последовательность действий**

1. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

Размер датасета - 9873 строк, 13 колонок,

Список колонок с типами данных:

Hours Studied int64

Previous Scores int64

Extracurricular Activities object

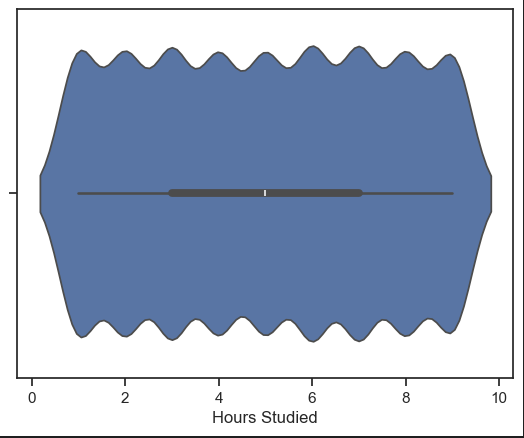
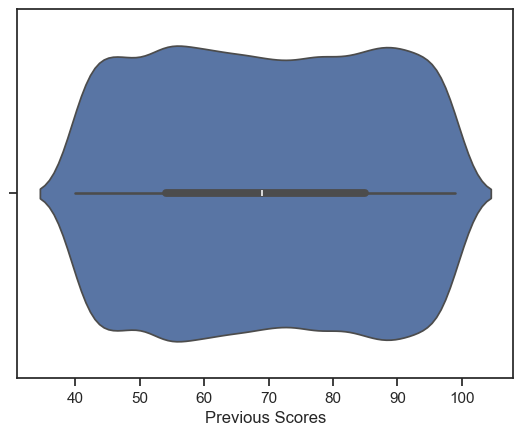
Sleep Hours int64

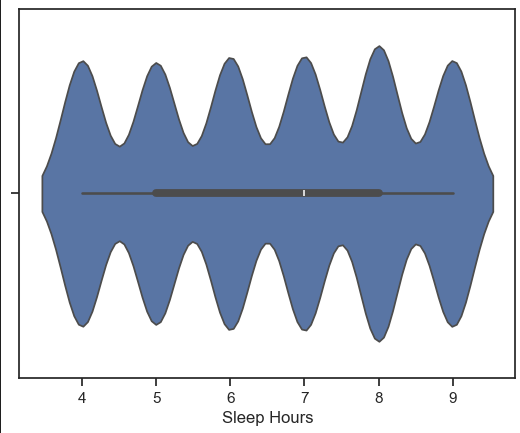
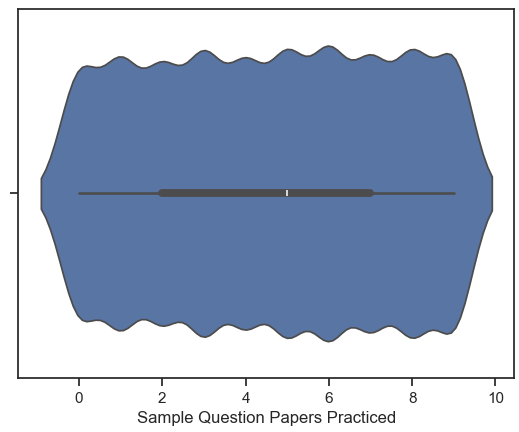
Sample Question Papers Practiced int64

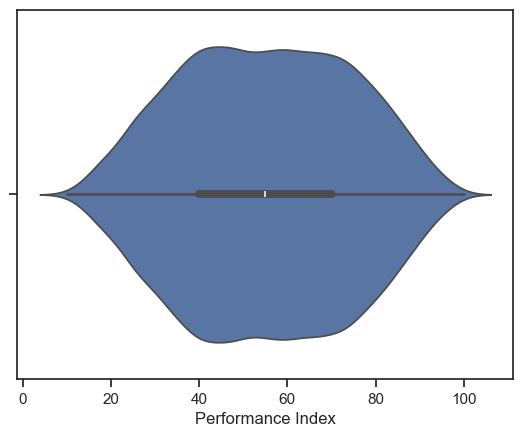
Performance Index float64,

Пустых значений не обнаружено, категориальные признаки закодируем с помощью LabelEncode.

Скрипичные диаграммы:

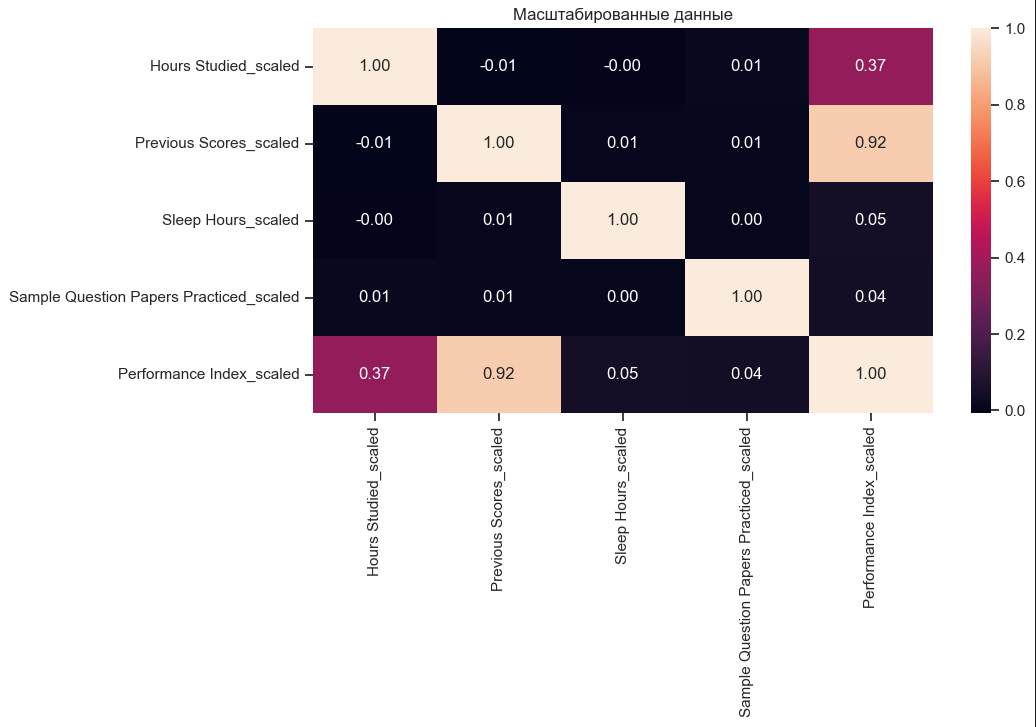


1. Подготовка Данных: В процессе подготовки данных были выполнены следующие действия:
   * Масштабирование числовых признаков, чтобы обеспечить равномерное распространение значений и улучшить стабильность обучения моделей.
2. Выбор Признаков: Был проведен анализ корреляции между признаками и целевой переменной "Performance Index" для определения наиболее значимых признаков, которые будут использоваться для построения моделей.

На основе корреляционной матрицы можно сделать следующие выводы:

- Корреляционные матрицы для исходных и масштабированных данных совпадают.

- Целевой признак классификации "Performance Index" наиболее сильно коррелирует с Previous Score (0.92) и Hourse Studied (0.37). И они не коррелируют между собой

- Большие по модулю значения коэффициентов корреляции свидетельствуют о значимой корреляции между исходными признаками и целевым признаком. На основании корреляционной матрицы можно сделать вывод о том, что данные позволяют построить модель машинного обучения. 

1. Выбор метрик для последующей оценки качества моделей.

Mean absolute error - средняя абсолютная ошибка.

Mean squared error - средняя квадратичная ошибка

Метрика R2

MAE и MSE взяты как как важна направленность ошибки, R2 так как важно, насколько хорошо модель объясняет вариацию в данных

1. Выбор наиболее подходящих моделей для решения задачи регрессии.

* Линейная регрессия
* Метод ближайших соседей
* Метод опорных векторов
* Решающее дерево
* Случайный лес
* Градиентный бустинг

1. Формирование обучающей и тестовой выборок на основе исходного набора данных.
2. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров.

LR MAE=0.026, MSE=0.001, R2=0.978

KNN\_5 MAE=0.022, MSE=0.001, R2=0.984

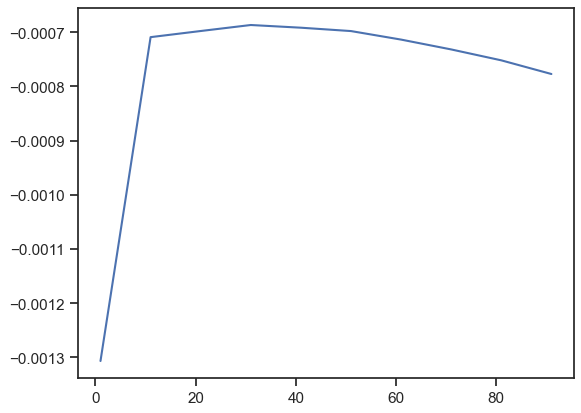
SVR MAE=0.024, MSE=0.001, R2=0.98

Tree MAE=0.021, MSE=0.001, R2=0.985

RF MAE=0.021, MSE=0.001, R2=0.985

GB MAE=0.02, MSE=0.001, R2=0.985

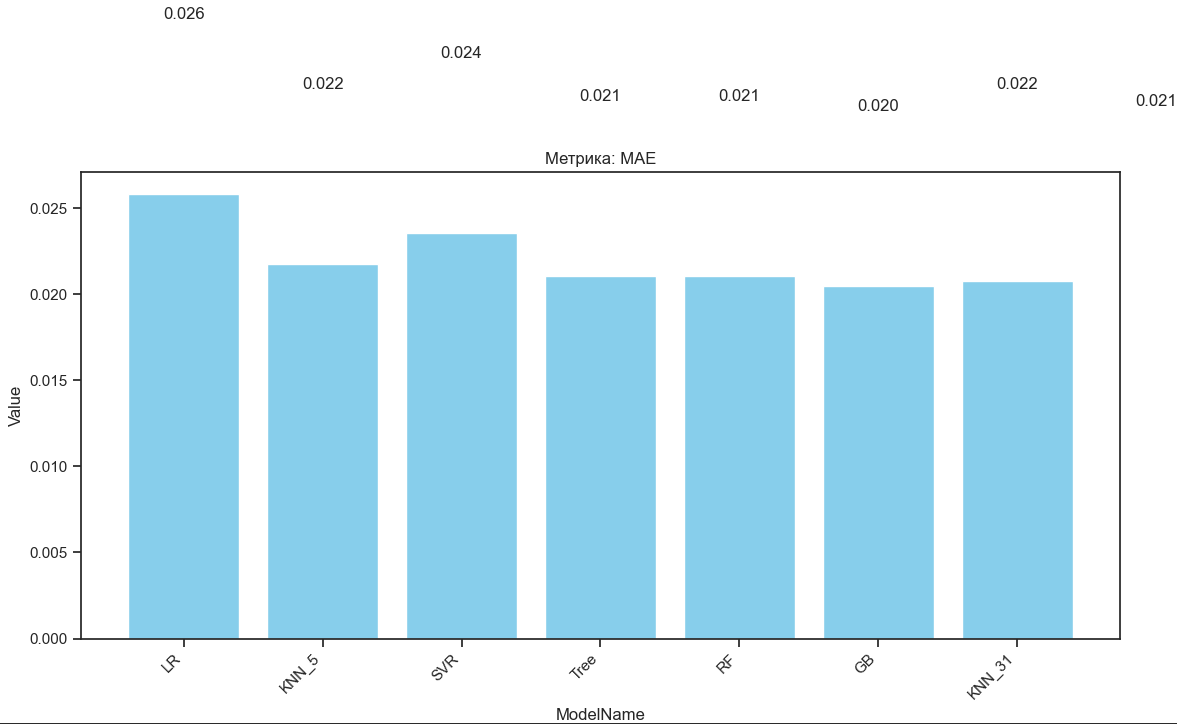
Подбор гиперпараметров для выбранных моделей. Анализ Результатов

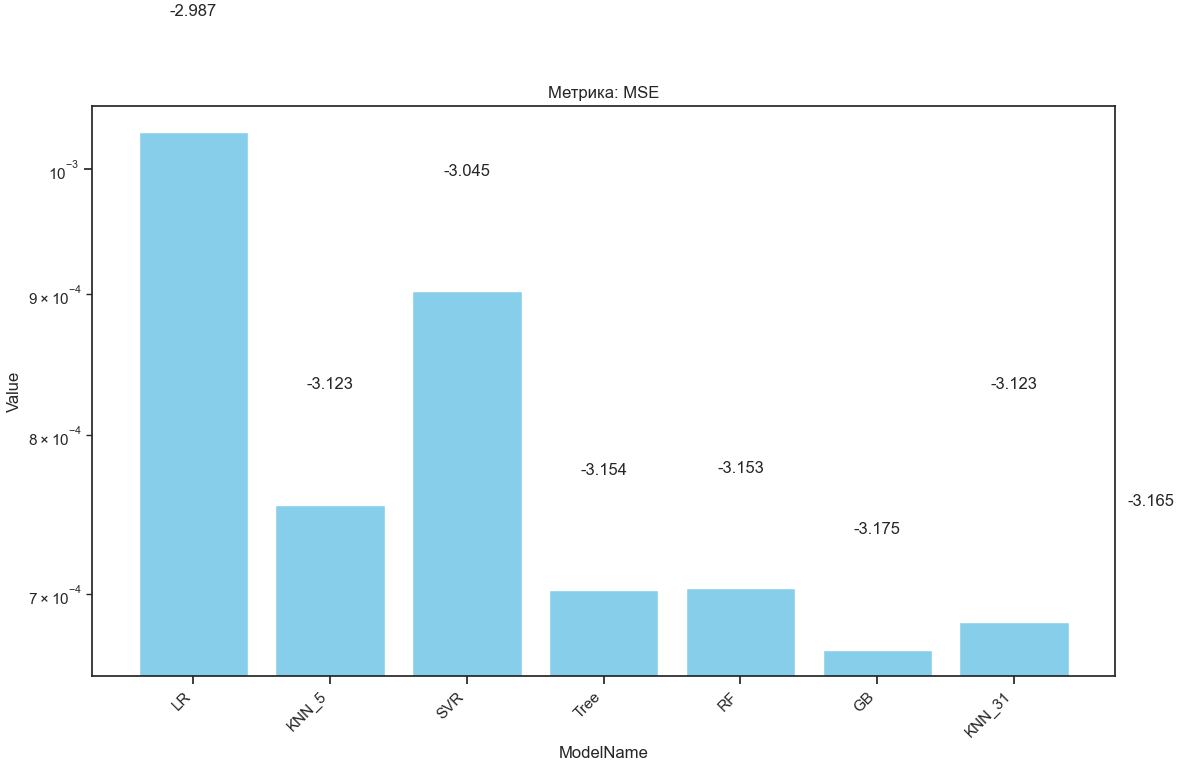


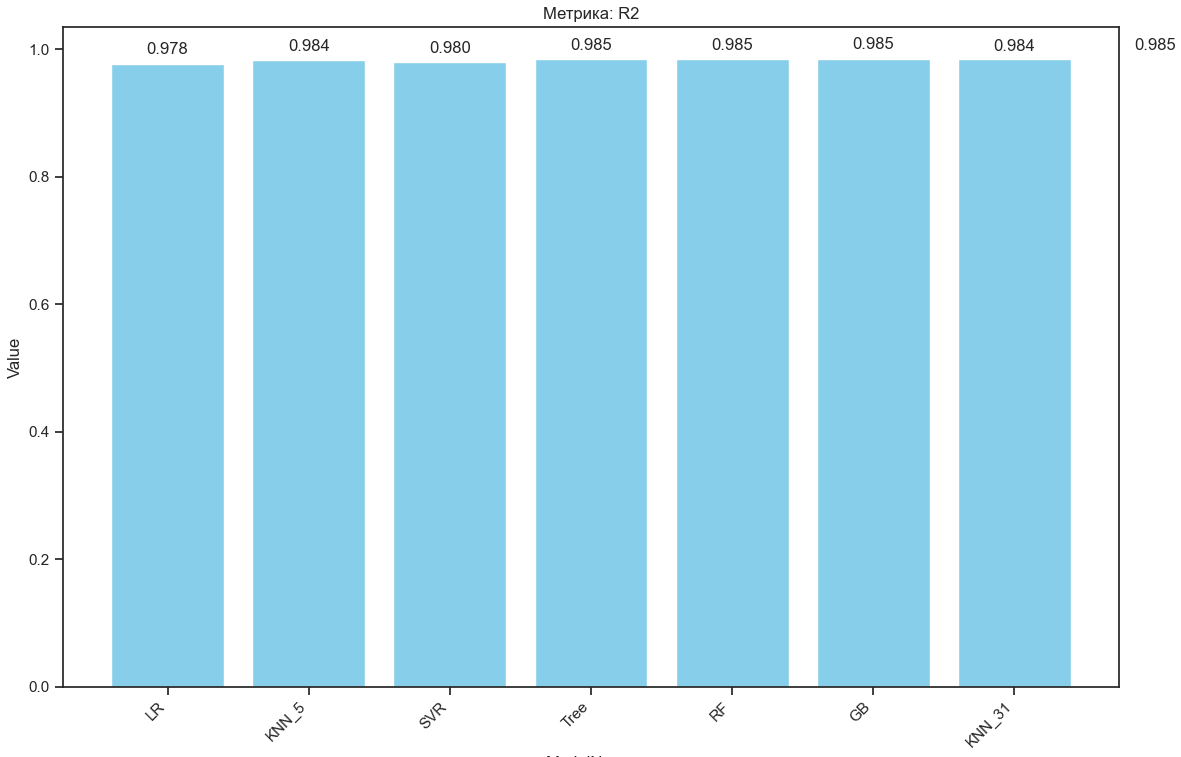
1. Сравнение значений базового и с гиперпараметром
2. KNN\_5 MAE=0.022, MSE=0.001, R2=0.984
3. KNN\_31 MAE=0.021, MSE=0.001, R2=0.985

Итог: гиперпараметры помогли незначительно улучшить результаты метрик

1. Формулирование Выводов.







По графикам видно, что лучше всего себя показал градиентный бустинг.

### Заключение

В ходе данной работы проведено всестороннее исследование для прогнозирования продуктивности студентов с использованием набора данных с Kaggle. Разработаны и оценены различные модели машинного обучения, включая как базовые, так и оптимизированные с применением подбора гиперпараметров.

Результаты показали, что подбор гиперпараметров незначительно улучшает точность прогнозирования, делая модели более точными и надежными. Особенно эффективно себя показала модель линейной регрессии.

Таким образом, созданные модели машинного обучения и проведенный анализ позволяют не только прогнозировать результативность студентов, но и выявлять ключевые факторы, которые следует учитывать при подготовке к конкурсу. Это может стать ценным инструментом как для мтудентов, так и для образовательных учреждений.

В будущем, расширение анализа за счет дополнительных признаков и использование более продвинутых методов машинного обучения может еще больше повысить точность прогнозирования и дать более глубокое понимание механизмов принятия решений в процессе обучения.