|  |  |
| --- | --- |
|  | **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.А. Надыршина\_\_\_**

(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\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Надыршина Алёна Альбертовна\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(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. А. Надыршина

(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 цене алмазов, предоставленного Kaggle. Этот набор данных включает информацию о количестве карат, качестве огранки, цвете, цене, чистоте, длине, ширине и высоте. Целью работы является разработка и оценка моделей машинного обучения для прогнозирования цены алмазов.

Модели машинного обучения могут предложить ценный инструмент для анализа сложных взаимосвязей между различными признаками и их влиянием на цену алмазов.

В ходе работы будет детальный анализ данных, включая предварительный исследовательский анализ, выбор и кодирование признаков, корреляционный анализ и подбор гиперпараметров для моделей машинного обучения. Также будет сравнение различных моделей, чтобы определить, какая из них наиболее эффективна в прогнозировании цены.

### Основная часть

**Постановка Задачи**

Цель этой работы — разработать и оценить модели машинного обучения для прогнозирования цены алмазов, используя набор данных с Kaggle. В качестве целевой переменной для задачи регрессии выбрана цена("price").

Датасет состоит из одного файла diamonds со следующими колонками

Carat – количество карат

Cut – качество огранки

Color – цвет

Clarity – чистота

Depth – глубина в процентах

Table – ширина вершины ромба относительно самой широкой точки

Price – цена

X - длина

Y - ширина

Z - высота

**Последовательность действий**

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

Список колонок с типами данных:

carat float64

cut object

color object

clarity object

depth float64

table float64

price int64

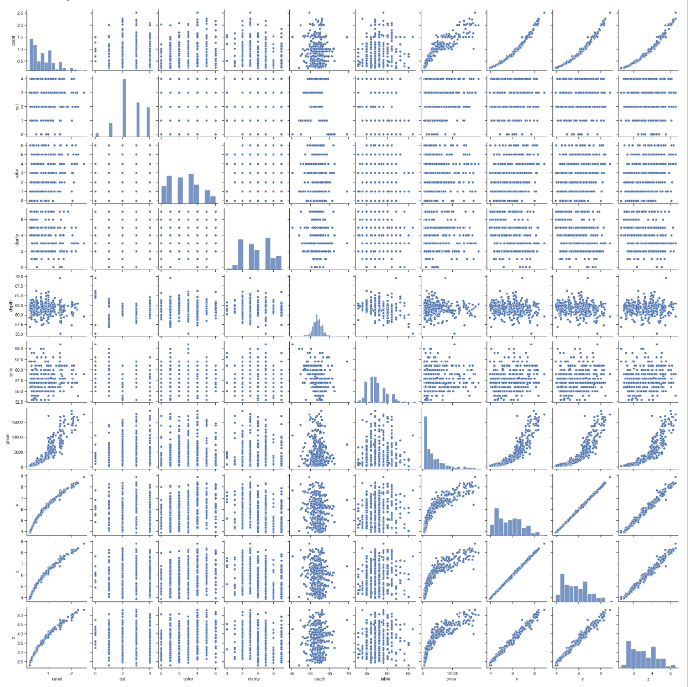
x float64

y float64

z float64

Пустых значений не обнаружено, признаки cut, color, clarity были закодированы.

Парные диаграммы:

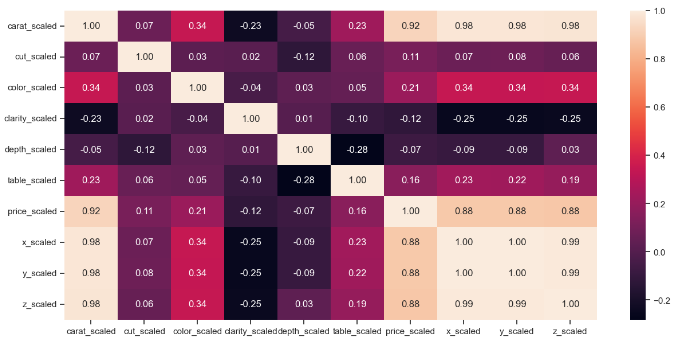


1. Подготовка Данных: В процессе подготовки данных были выполнены следующие действия:
   * Масштабирование числовых признаков, чтобы обеспечить равномерное распространение значений и улучшить стабильность обучения моделей.
2. Выбор Признаков: Был проведен анализ корреляции между признаками и целевой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

На основе корреляционной матрицы можно сделать следующие выводы:

- Целевой признак классификации "price" наиболее сильно коррелирует с carat(0.92), x(0.88), y(0.88), z(0.88).

- 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

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=0.052, MSE=0.01, R2=0.811

KNN\_5 MAE=0.055, MSE=0.011, R2=0.784

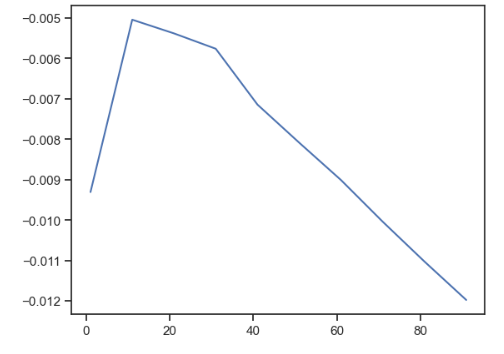
SVR MAE=0.063, MSE=0.01, R2=0.8

Tree MAE=0.059, MSE=0.013, R2=0.737

RF MAE=0.052, MSE=0.01, R2=0.795

GB MAE=0.054, MSE=0.012, R2=0.766

1. Подбор гиперпараметров для выбранных моделей. Анализ Результатов



1. Сравнение значений базового и с гиперпараметром

KNN\_5 MAE=0.055, MSE=0.011, R2=0.784

SVR MAE=0.063, MSE=0.01, R2=0.8

Tree MAE=0.059, MSE=0.013, R2=0.737

RF MAE=0.052, MSE=0.01, R2=0.795

GB MAE=0.054, MSE=0.012, R2=0.766

Best Model (KNN) MAE=0.053, MSE=0.01, R2=0.793

Best Model (Random Forest) MAE=0.051, MSE=0.011, R2=0.79

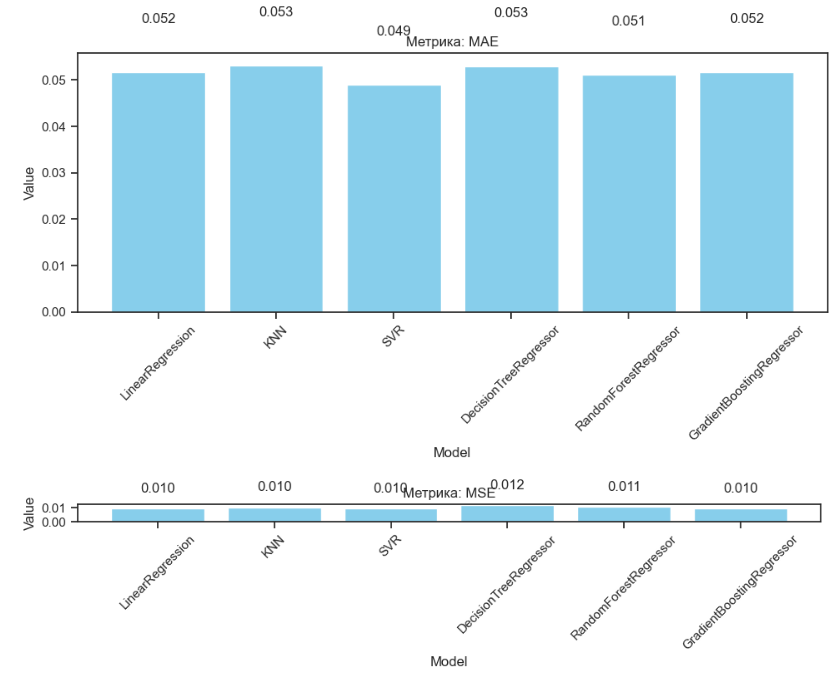
Best Model (SVR) MAE=0.049, MSE=0.01, R2=0.811

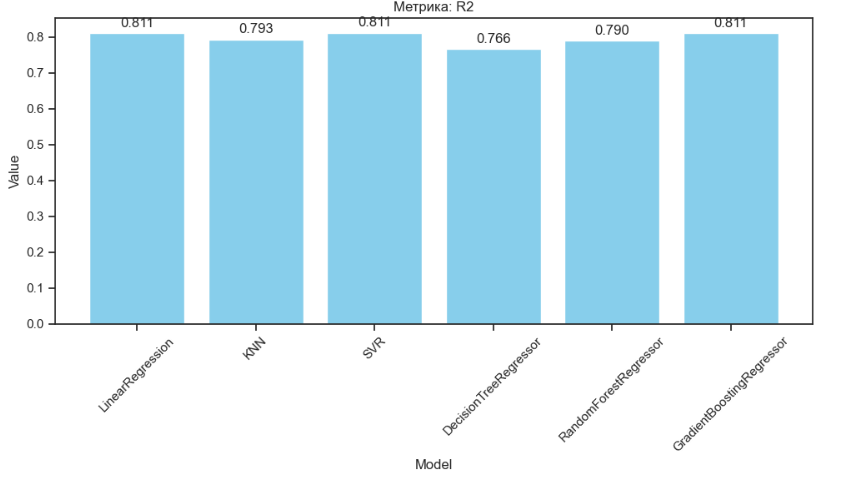
Best Model (DecisionTreeRegressor) MAE=0.053, MSE=0.012, R2=0.766

Best Model (GradientBoostingRegressor) MAE=0.06, MSE=0.011, R2=0.773

Итог: гиперпараметры помогли незначительно улучшить результаты метрик

1. Формулирование Выводов.





Лучше всего себя показала модель опорных векторов.

### Заключение

В ходе данной работы проведено всестороннее исследование для прогнозирования содержания pH в вине 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 градиентный бустинг.