МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования

«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ»

КАФЕДРА №  43

ОТЧЁТ

ЗАЩИЩЁН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

### ст.пр.                                   Поляк М.Д.

должность, уч. Степень, звание   подпись, дата           инициалы, фамилия

ОТЧЁТ О ЛАБОРАТОРНОЙ РАБОТЕ №2.

Построение моделей линейной регрессии.

по курсу: Основы машинного обучения

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ ГР. 4136                                                                                Бобрович Н. С.

                                                                         подпись, дата                      инициалы, фамилия

Санкт-Петербург 2024

1. **Цель работы:**

Получение навыков решения задачи регрессионного анализа и оптимизации функций методом градиентного спуска.

1. **Индивидуальное задание:**

Задание 1

Откройте Jupyter-ноутбук [regression\_assignment.ipynb](https://github.com/suai-mlb-2024/mlb-task2-OskolockKoli/blob/main/regression_assignment.ipynb) в этом репозитории. Скопируйте путь в адресной строке браузера. Перейдите в [Google Colab](https://colab.research.google.com/), в меню выберите "Файл" -> "Открыть ноутбук", в открывшемся окне слева выбрать "GitHub", затем:

вставить в поле для поиска скопированный URL;

поставить галочку "Показывать личные хранилища" ("Include private repos");

и нажать на иконку с лупой. При необходимости разрешить Colab доступ к аккаунту GitHub, если откроется новое окно с таким приглашением. Среди результатов поиска выбрать regression\_assignment.ipynb и приступить к выполнению задания.

1. **Ход работы:**

Задание 1:

**Фрагмент кода из задания 1.**

### BEGIN YOUR CODE

# Загрузка датасетаdf = pd.read\_csv('forestfires.csv')

# Каков размер датасета? (в Мб)memory\_usage\_in\_bytes = sys.getsizeof(df)memory\_usage\_in\_mb = memory\_usage\_in\_bytes / (1024 \* 1024)print(f'\nРазмер датасета в памяти: {memory\_usage\_in\_mb:.2f} МБ')

#Сколько в нем записей (объектов)?n\_rows = len(df)print(f'\nКоличество записей (объектов): {n\_rows}')

#Сколько признаков (фич) у объектов в датасете?n\_columns = df.shape[1]print(f'\nКоличество признаков (фич): {n\_columns}')

#Есть ли категориальные данные? Какие?categorical\_cols = ['month', 'day']for col in categorical\_cols:

print(f'\n{col}: {df[col].unique()}')

#Есть ли пропущенные значения?has\_missing\_values = df.isnull().values.any()if has\_missing\_values:

missing\_values\_count = df.isna().sum()

print(f'\nПропущенные значения: \n{missing\_values\_count}')else:

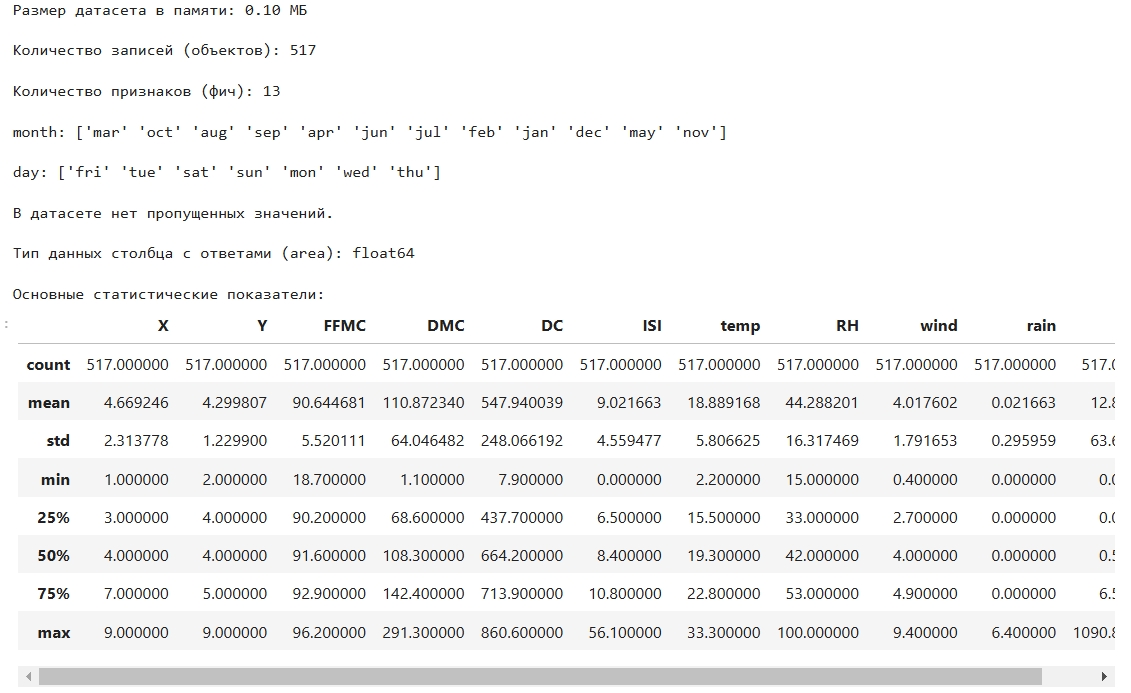
print('\nВ датасете нет пропущенных значений.')

#Есть ли в датасете столбец с ответами (target)? Какой у него тип данных?target\_col = 'area'print(f'\nТип данных столбца с ответами ({target\_col}): {df[target\_col].dtype}')

print('\nОсновные статистические показатели:')df.describe()

### END YOUR CODE

**Результат выполнения кода из задания 1.**



**Пояснения к коду из задания 1.**

Датасет содержит информацию о лесных пожарах, включающую различные метеорологические параметры и характеристики пожарной активности.

Размер датасета в памяти составляет 0.10 МБ. В датасете представлено 517 записей, каждая из которых включает 13 признаков. Среди них два категориальных признака: month (месяц) и day (день недели). Пропущенные значения отсутствуют. Целевая переменная area имеет тип данных float64 и представляет собой площадь лесного пожара. Статистический анализ показывает, что средняя площадь пожара составляет 12.85 га, однако диапазон значений варьируется от 0 до 1090.84 га, что свидетельствует о большом разбросе данных. Таким образом, регрессионная модель, построенная на этом датасете, будет предсказывать площадь лесного пожара на основе предоставленных признаков, таких как индекс пожарной опасности, температура, влажность и скорость ветра.

1. **Выводы:**

Получил навыки решения задачи регрессионного анализа и оптимизации функций методом градиентного спуска.