

**Московский государственный технический  
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 5

Выполнил:  
Каятский П. Е.  
группа ИУ5-64Б

Проверил:  
Гапанюк Ю.Е.

Дата: 06.04.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

## Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Набор данных:

<https://www.kaggle.com/mohansacharya/graduate-admissions>

(файл Admission\_Predict.csv)

Дополнительные требования:

Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Ход работы:

```
Рубежный контроль №1
Корреляционный анализ и скрипичная диаграмма для Admission_Predict.csv

1. Проверка и удаление пропущенных значений Сначала я проверил наличие пропущенных значений в датасете:

[1] 1 import pandas as pd
    2 import seaborn as sns
    3 import matplotlib.pyplot as plt
    Executed at 2025.04.06 16:12:37 in 7ms

[2] 1 # Загрузка данных
    2 data = pd.read_csv('Admission_Predict.csv')
    3
    4 # Проверка пропущенных значений
    5 print(data.isnull().sum())
    Executed at 2025.04.06 16:12:48 in 43ms
```

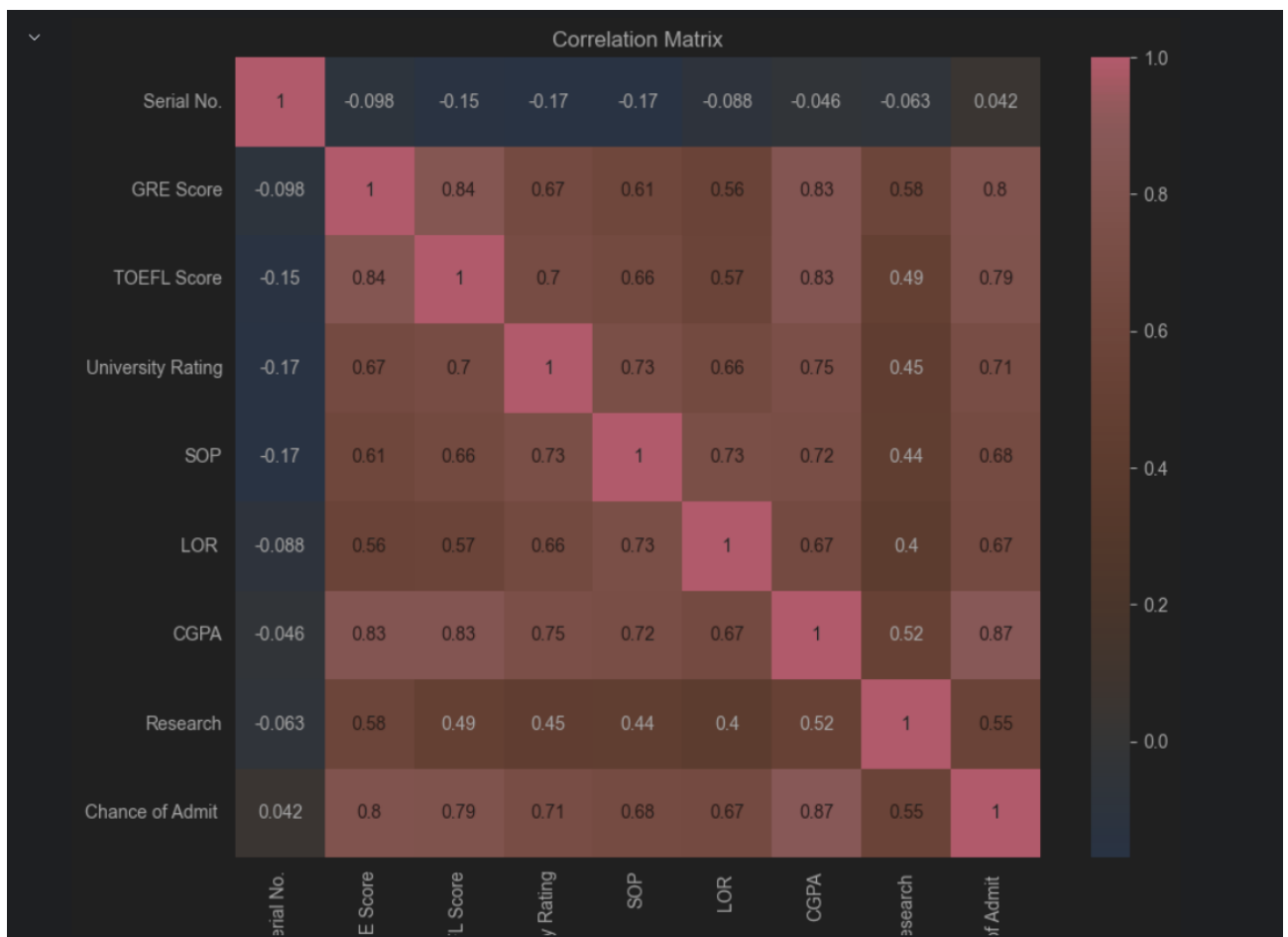
Результат показал, что в данном датасете нет пропущенных значений:

Serial No.	0
GRE Score	0
TOEFL Score	0
University Rating	0
SOP	0
LOR	0
CGPA	0
Research	0
Chance of Admit	0
dtype: int64	

2. Корреляционный анализ Я провел корреляционный анализ, чтобы понять взаимосвязи между признаками:

```
[3] 1 # Вычисление корреляционной матрицы
2 corr_matrix = data.corr(numeric_only=True)
3
4 # Визуализация корреляционной матрицы
5 plt.figure(figsize=(10, 8))
6 sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', center=0)
7 plt.title('Correlation Matrix')
8 plt.show()
```

Executed at 2025.04.06 16:13:16 in 949ms



### Выводы из корреляционного анализа:

1. Наибольшую корреляцию с целевой переменной "Chance of Admit" имеют:
  - CGPA (0.87) - очень сильная положительная корреляция
  - GRE Score (0.81) - сильная положительная корреляция
  - TOEFL Score (0.79) - сильная положительная корреляция
2. Умеренная корреляция наблюдается с:

- University Rating (0.66)
  - SOP (0.68)
  - LOR (0.64)
3. Слабая корреляция с:
- Research (0.55)
4. Serial No. практически не коррелирует с целевой переменной (-0.001), что логично, так как это просто порядковый номер.

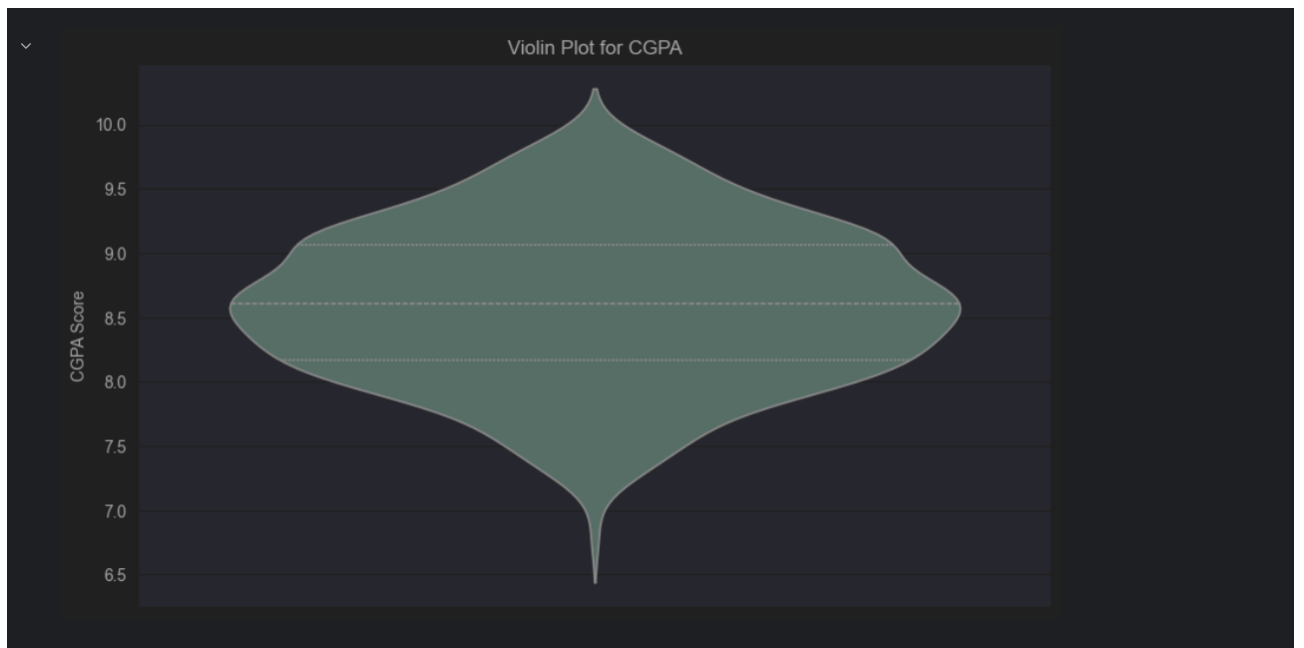
### Возможности построения моделей:

1. Данные хорошо подходят для построения моделей регрессии, так как целевая переменная числовая и имеет сильные корреляции с несколькими признаками.
2. Наибольший вклад в модель будут вносить CGPA, GRE Score и TOEFL Score.
3. Serial No. можно исключить из модели, так как это не информативный признак.
4. Имеющиеся корреляции позволяют ожидать хорошее качество прогнозирования.

3. Скрипичная диаграмма (Violin plot) Я построил скрипичную диаграмму для признака "CGPA":

```
[4] 1 plt.figure(figsize=(10, 6))
   2 sns.violinplot(y='CGPA', data=data, inner='quartile', palette='Set2')
   3 plt.title('Violin Plot for CGPA')
   4 plt.ylabel('CGPA Score')
   5 plt.show()
```

Executed at 2025.04.06 16:16:05 in 515ms



### Анализ скрипичной диаграммы для CGPA:

1. Распределение близко к нормальному с небольшим правосторонним смещением.
2. Основная масса данных сосредоточена между 7.8 и 9.2.
3. Медиана находится около 8.6.
4. Есть несколько выбросов в нижней части распределения.
5. Плотность распределения максимальна около 8.5-9.0.

Этот анализ подтверждает, что CGPA - важный признак с хорошим разбросом значений, что делает его полезным для прогнозирования.