Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 7

Выполнил:
Искорнев И. П.
группа ИУ5-61Б

Проверил:
Гапанюк Ю.Е.

Дата: 12.05.25

Подпись:

Дата:

Подпись:

Москва, 2025 г.

Задание:

Номер варианта: **7**

Номер задачи: **1**

Номер набора данных, указанного в задаче: 1 (https://www.kaggle.com/datasets/mohansacharya/graduate-admissions)

Для студентов группы ИУ5-61Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

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.

1. Введение

В рамках рубежного контроля была проведена работа с набором данных Admission Prediction. Целью работы являлось заполнение пропусков данных.

2. Описание исходных данных

В данном датасете содержатся данные о проценте зачисления в магистратуру и какие признаки влияют на процент поступления.

3. Ход выполнения:

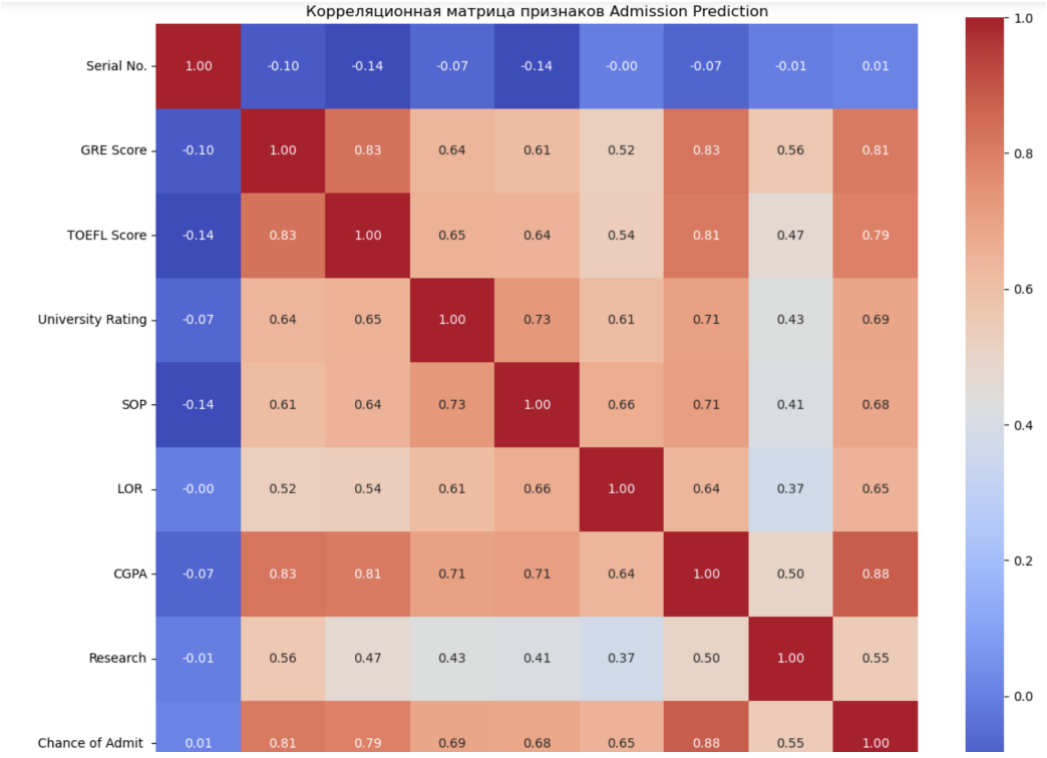
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# 1. Загрузка данных
df = pd.read_csv("Admission_Predict_Ver1.1.csv")

# 1.1 Изучение датасета
print("Первые 5 строк данных:")
print(df.head())
print("\nИнформация о данных:")
print(df.info())
print("\nОписательная статистика:")
print(df.describe())
```

```
Первые 5 строк данных:
   Serial No.  GRE Score  TOEFL Score  University Rating  SOP  LOR   CGPA  \
0           1        337          118                  4  4.5   4.5  9.65
1           2        324          107                  4  4.0   4.5  8.87
2           3        316          104                  3  3.0   3.5  8.00
3           4        322          110                  3  3.5   2.5  8.67
4           5        314          103                  2  2.0   3.0  8.21

   Research  Chance of Admit
0         1             0.92
1         1             0.76
2         1             0.72
3         1             0.80
4         0             0.65
```

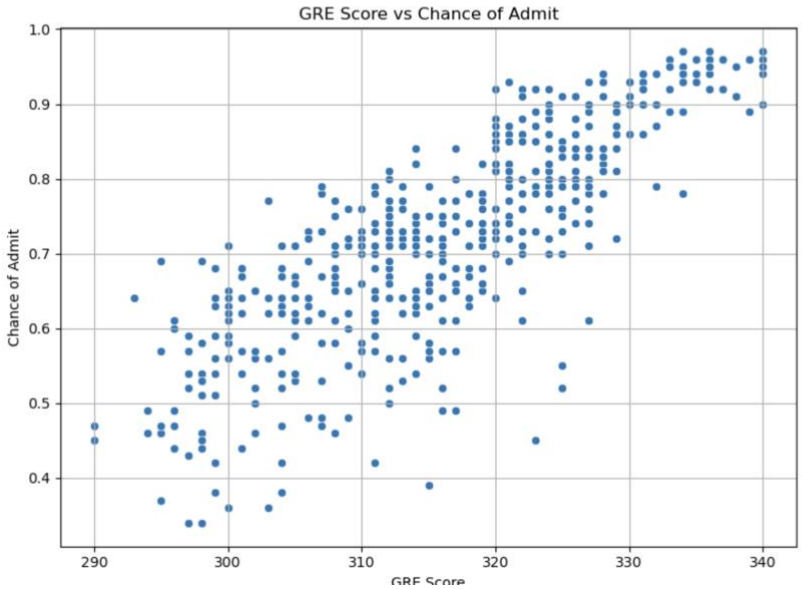
```
Информация о данных:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Serial No.         500 non-null    int64
 1   GRE Score          500 non-null    int64
 2   TOEFL Score        500 non-null    int64
 3   University Rating  500 non-null    int64
 4   SOP                500 non-null    float64
 5   LOR                500 non-null    float64
 6   CGPA               500 non-null    float64
 7   Research           500 non-null    int64
 8   Chance of Admit    500 non-null    float64
dtypes: float64(4), int64(5)
memory usage: 35.3 KB
None

Описательная статистика:
       Serial No.   GRE Score  TOEFL Score  University Rating         SOP  \
count  500.000000  500.000000   500.000000         500.000000  500.000000
mean   250.500000  316.472000   107.192000           3.114000    3.374000
std    144.481833   11.295148     6.081868           1.143512    0.991004
min      1.000000  290.000000    92.000000           1.000000    1.000000
25%    125.750000  308.000000   103.000000           2.000000    2.500000
50%    250.500000  317.000000   107.000000           3.000000    3.500000
75%    375.250000  325.000000   112.000000           4.000000    4.000000
max    500.000000  340.000000   120.000000           5.000000    5.000000

             LOR         CGPA    Research  Chance of Admit
count  500.00000  500.000000  500.000000         500.00000
mean     3.48400    8.576440    0.560000           0.72174
std      0.92545    0.604813    0.496884           0.14114
min      1.00000    6.800000    0.000000           0.34000
25%      3.00000    8.127500    0.000000           0.63000
50%      3.50000    8.560000    1.000000           0.72000
75%      4.00000    9.040000    1.000000           0.82000
max      5.00000    9.920000    1.000000           0.97000
```

```python
# 2. Проверка пропусков
print("Количество пропусков по колонкам:")
print(df.isnull().sum())

# 3. Удаление строк с пропущенными значениями
df_clean = df.dropna()
print("\nФорма набора после удаления пропусков:", df_clean.shape)

# 4. Корреляционный анализ
corr_matrix = df_clean.corr(numeric_only=True)

# 5. Визуализация корреляционной матрицы
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", square=True)
plt.title("Корреляционная матрица признаков Admission Prediction")
plt.tight_layout()
plt.show()
```

```
Количество пропусков по колонкам:
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

Форма набора после удаления пропусков: (500, 9)
```



```
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df_clean, x='GRE Score', y='Chance of Admit ')
plt.title("GRE Score vs Chance of Admit")
plt.xlabel("GRE Score")
plt.ylabel("Chance of Admit")
plt.grid(True)
plt.tight_layout()
plt.show()
```



1. Возможность построения моделей машинного обучения

- Признаки в наборе данных демонстрируют **существенную корреляцию** с целевой переменной **Chance of Admit**, что делает возможным применение регрессионных моделей или моделей на основе деревьев решений.
- Корреляционный анализ позволяет оценить, какие признаки **наиболее информативны** — чем выше модуль корреляции с **Chance of Admit**, тем выше их значимость при прогнозировании шанса поступления.

2. Наиболее важные признаки (по корреляции с `Chance of Admit`)

Признак	Корреляция с `Chance of Admit`	Влияние на модель
`CGPA`	**+0.88**	Сильное положительное влияние: больше GPA — выше шанс
`GRE`	**+0.81**	Сильное положительное: высокий балл GRE улучшает шанс
`TOEFL Score`	**+0.79**	Сильное положительное: хороший результат TOEFL увеличивает шансы
`Research`	**+0.55**	Умеренное влияние: Наличие исследовательского опыта даёт преимущество

3. Взаимные корреляции (мультиколлинеарность)

- Некоторые признаки, такие как **`GRE Score`**, **`TOEFL Score`** и **`CGPA`**, имеют высокую взаимную корреляцию 0.83. Это может привести к мультиколлинеарности в линейной регрессии.

Итоговый вывод:

Данный набор данных подходит для построения **моделей регрессии**, в частности — **линейной или градиентного бустинга**., а признаки `CGPA`, `GRE` и `TOEFL Score` имеют наибольший вклад в целевую переменную, однако необходимо учитывать взаимные зависимости между признаками и избегать их дублирования.

4. Выводы

В ходе работы были успешно выполнены задачи по обработке данных, заполнению пропусков, кодированию категориальных данных и визуализации диаграмм рассеяния.