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

Факультет «Информатика и системы управления»
Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»
Отчет по рубежному контролю №1
«Технологии разведочного анализа и обработки данных»
Вариант №6


Выполнил:
студент группы ИУ5-61Б
Зелинский Даниил
Михайлович

Подпись:_____

Дата:_____

Проверил:
преподаватель каф. ИУ5
Гапанюк Юрий
Евгеньевич

Подпись:_____

Дата:_____


Москва, 2023 г.

# Выполнение работы

Для выполнения задачи проведения корреляционного анализа данных был использован набор данных Admission_Predict_Ver1.1.

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
df = pd.read_csv(r'Admission_Predict_Ver1.1.csv')
df.head()
```

Out[4]:

| | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 337 | 118 | 4 | 4.5 | 4.5 | 9.65 | 1 | 0.92 |
| **1** | 2 | 324 | 107 | 4 | 4.0 | 4.5 | 8.87 | 1 | 0.76 |
| **2** | 3 | 316 | 104 | 3 | 3.0 | 3.5 | 8.00 | 1 | 0.72 |
| 3 | 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.80 |
| **4** | 5 | 314 | 103 | 2 | 2.0 | 3.0 | 8.21 | 0 | 0.65 |

Типы данных всех полей являются числовыми.

In [5]:
```
df.info()
```

```
<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
```

В наборе данных отсутствуют пропуски и дубликаты.

```
In [7]: df.isna().sum()

Out[7]: 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
```

```
In [35]: df.duplicated().sum()

Out[35]: 0
```

Из датафрейма были выделены 7 нецелевых признаков и 1 целевой – рейтинг университета. В процессе преобразования датафрейма целевой признак был вынесен в последний столбец.

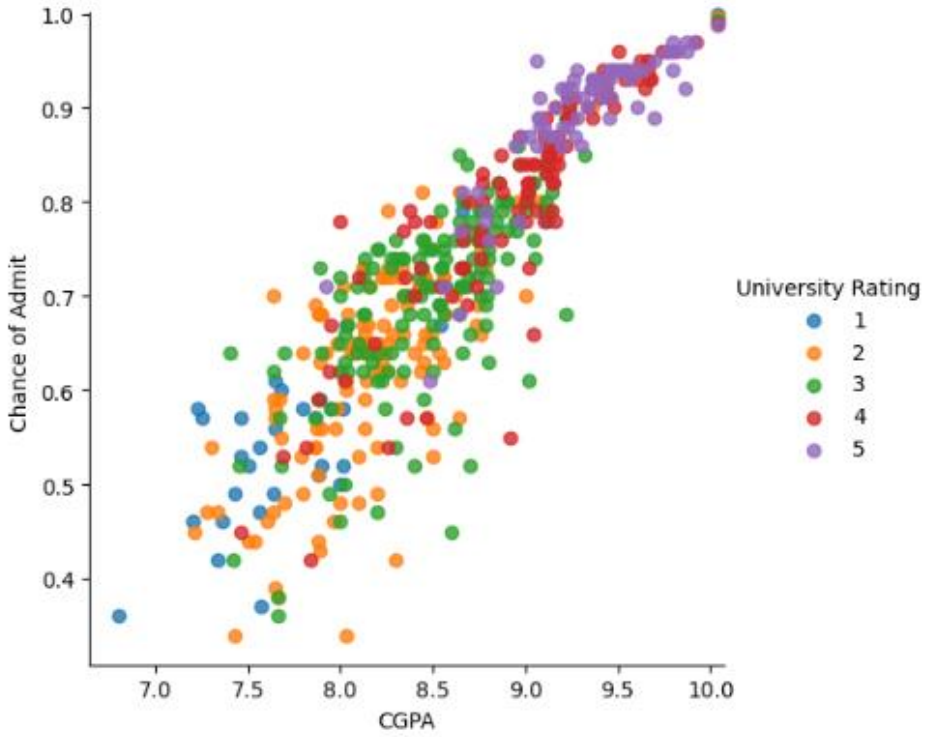
```
In [75]: # разделение на объекты-признаки и целевой признак
df_x = pd.DataFrame(df.iloc[:, df.columns != 'University Rating'].values,
                    columns=[column for column in df.columns if column!='University Rating'] )
df_y = pd.DataFrame(df.iloc[:, -6].values, columns=['University Rating'])
```

```
In [111]: _df=pd.concat([df_x, df_y.reindex(df_x.index)], axis=1)

In [112]: _df.head()

Out[112]:
```

| | Serial No. | GRE Score | TOEFL Score | SOP | LOR | CGPA | Research | Chance of Admit | University Rating |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1.0 | 337.0 | 118.0 | 4.5 | 4.5 | 9.65 | 1.0 | 0.92 | 4 |
| **1** | 2.0 | 324.0 | 107.0 | 4.0 | 4.5 | 8.87 | 1.0 | 0.76 | 4 |
| **2** | 3.0 | 316.0 | 104.0 | 3.0 | 3.5 | 8.00 | 1.0 | 0.72 | 3 |
| **3** | 4.0 | 322.0 | 110.0 | 3.5 | 2.5 | 8.67 | 1.0 | 0.80 | 3 |
| **4** | 5.0 | 314.0 | 103.0 | 2.0 | 3.0 | 8.21 | 0.0 | 0.65 | 2 |

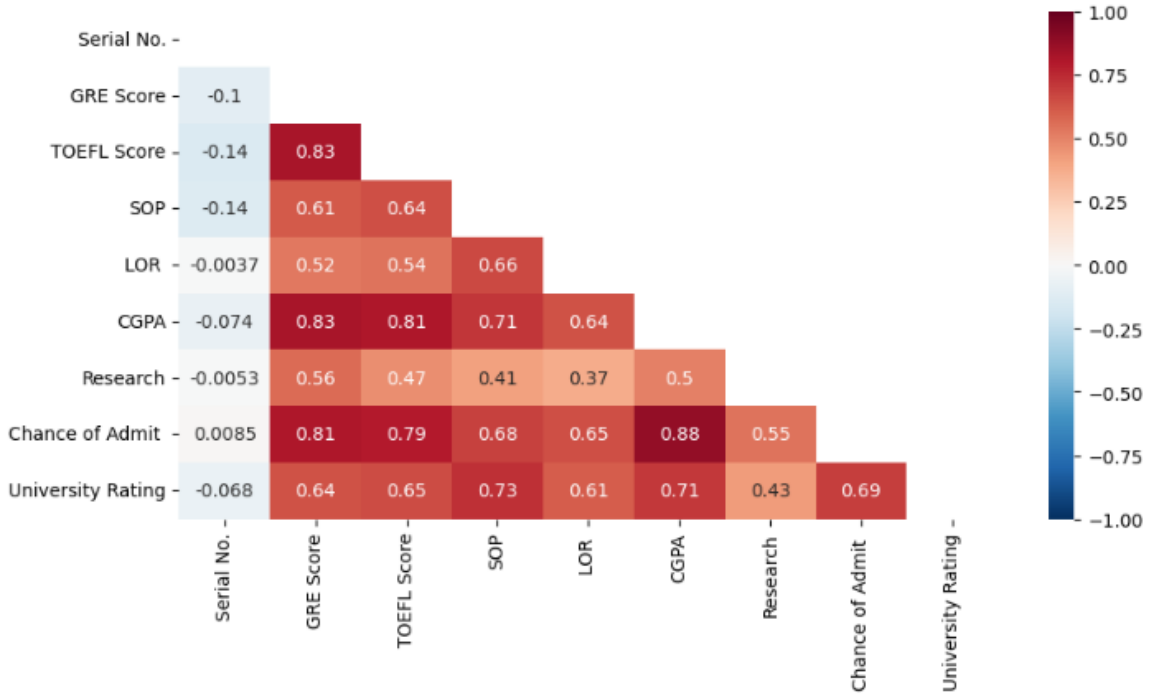
Для столбцов “Chance of Admit” и “CGPA” был построен график “Диаграмма рассеяния”.

```
In [13]: sns.lmplot(x=_df.columns[5], y=_df.columns[7], data=_df, fit_reg=False, hue='University Rating')
         plt.legend(loc='upper right', fontsize=0)
         plt.show()
```




Для визуализации корреляционной матрицы была использована “тепловая карта”.

```
In [14]: plt.figure(figsize=(10,5))
         m=np.triu(np.ones_like(_df.corr(), dtype=bool))
         sns.heatmap(_df.corr(), mask=m, annot=True, vmin=-1.0, vmax=1.0, center=0, cmap='RdBu_r')
Out[14]: <AxesSubplot: >
```



С целевым признаком “University Rating” наиболее коррелируют признаки “SOP” (0,73), “CGPA” (0,71), “Chance of Admit” (0,69), “TOEFL Score” (0,65), “GRE Score” (0,64), “LOR” (0,61). При построении модели машинного обучения перечисленные признаки будут наиболее информативными.

Целевой признак коррелирует отчасти с признаком “Research” (0,43), который также можно применять в процессе обучения модели. Признак “Serial No.” не коррелирует не только с целевым признаком (-0,068), но и со всеми остальными ввиду того, что предназначен для нумерации записей в наборе данных. Такой признак не привнесёт пользы в обучение моделей, и его следует изъять.

Стоит отметить корреляцию признаков “Chance of Admit” и “CGPA” (0,88) – самую высокую в датасете. Ввиду того, что первый признак – шанс поступления – зависит от среднего балла в дипломе (CGPA), из признаков, используемых для обучения модели, следует убрать “Chance of Admit”, несмотря на его корреляцию с целевым признаком.

Наконец, можно построить модель машинного обучения на основе признаков “SOP”, “CGPA”, “TOEFL Score”, “GRE Score” и “LOR”. Обученные модели позволят бакалаврам оценить свои возможности для поступления на магистратуру.