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 «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных»

Вариант №5

Выполнил:

студент группы ИУ5-63Б

Журмилов В.Д.

Преподаватель:

Гапанюк Ю. Е.

2024 г.

Задание:

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 машинногообучения и о возможном вкладе признаков в модель.

Для студентов групп ИУ5-63Б, ИУ5Ц-83Б - для произвольной колонкиданных построить график "Ящик с усами (boxplot)".

Набор данных:

https://www.kaggle.com/datasets/mohansacharya/graduate-admissions (файл Admission_Predict.csv)

Решение:

Подключим все необходимые библиотеки, загрузим набор данных ипроверим, что все успешно подключилось:

```
In [10]: import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
         plt.rcParams.update({'figure.max_open_warning': 0})

In [11]: data = pd.read_csv('Admission_Predict.csv')

In [13]: data.head()
```

Out[13]:

	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

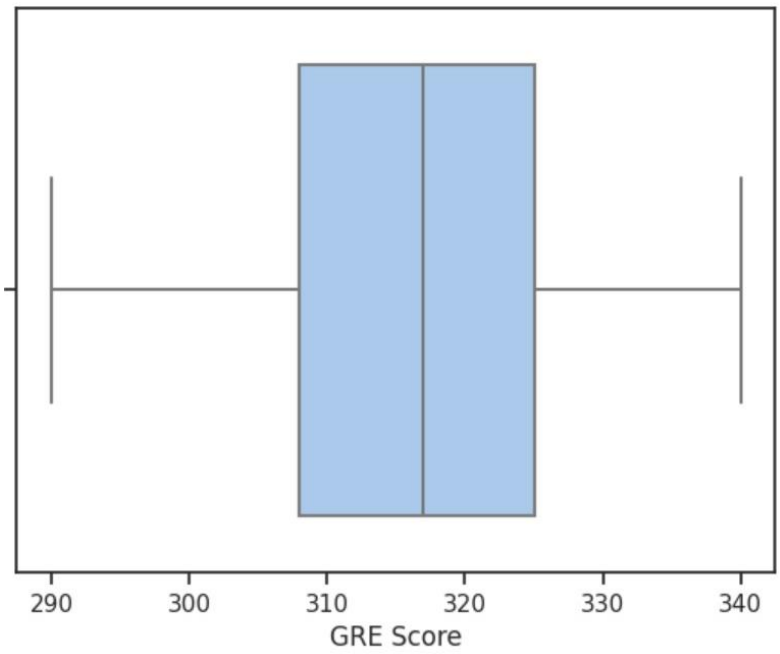
Проверяем набор данных на наличие пропусков:

```
In [14]: data.isna().sum()
Out[14]: 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
```

Пропуски отсутствуют.

По колонке “GRE Score” сделаем boxplot (Ящик с усами):

```
In [15]: sns.set_theme(style="ticks", palette="pastel")
         sns.boxplot(x=data["GRE Score"])
```

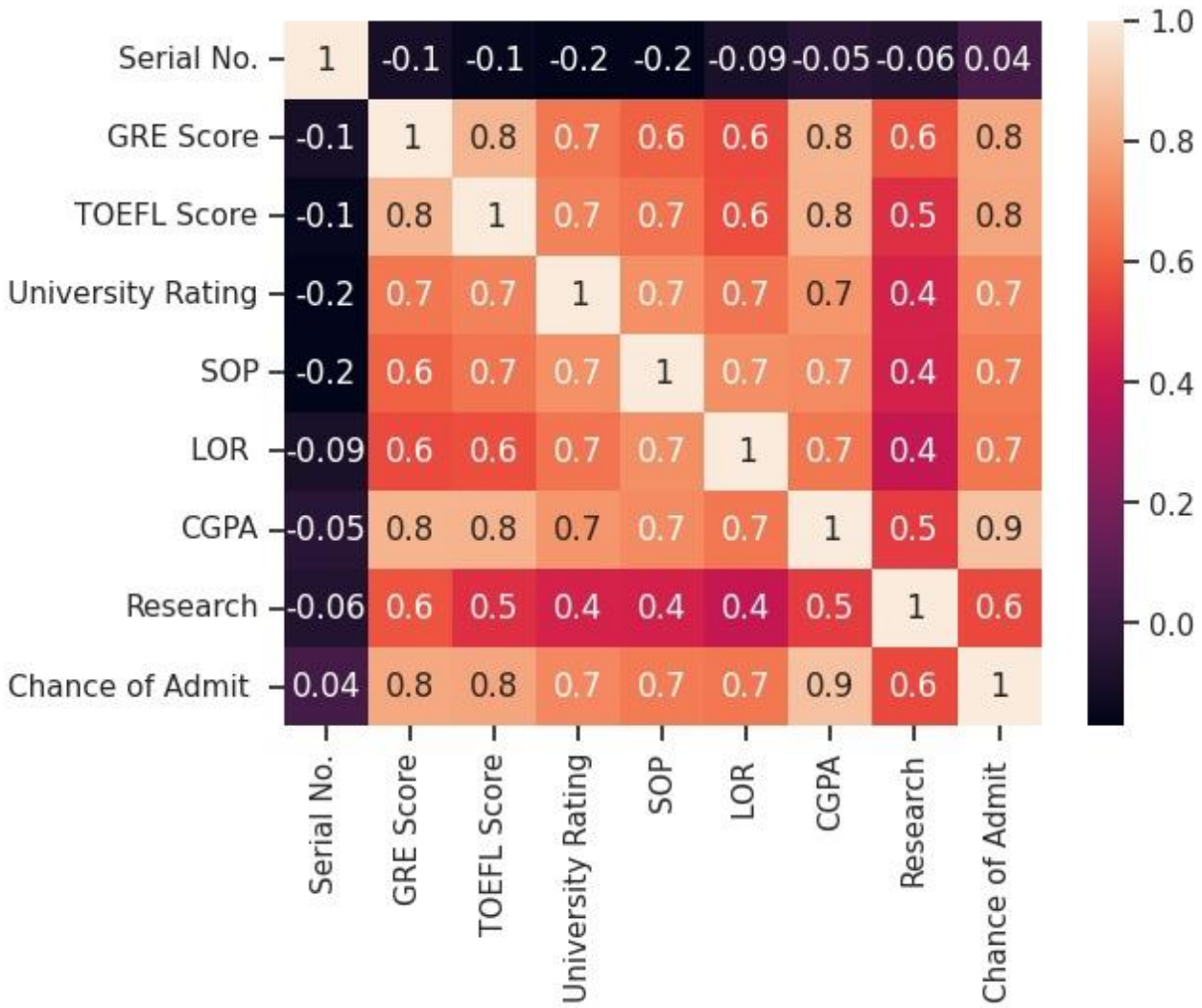


Данный график позволяет увидеть минимальные/максимальные значения,медиану, верхний и нижний квартили.

Перейдем к корреляционному анализу:

С помощью библиотеки seaborn создадим тепловую карту и изучимполучившиеся значения (ниже)

```
In [16]: sns.heatmap(data.corr(), annot = True, fmt='.1g')
```


Serial No.
GRE Score
TOEFL Score
University Rating
SOP
LOR
CGPA
Research
Chance of Admit
1 -0.1 -0.1 -0.2 -0.2 -0.09 -0.05 -0.06 0.04
-0.1 1 0.8 0.7 0.6 0.6 0.8 0.6 0.8
-0.1 0.8 1 0.7 0.7 0.6 0.8 0.5 0.8
-0.2 0.7 0.7 1 0.7 0.7 0.7 0.4 0.7
-0.2 0.6 0.7 0.7 1 0.7 0.7 0.4 0.7
-0.09 0.6 0.6 0.7 0.7 1 0.7 0.4 0.7
-0.05 0.8 0.8 0.7 0.7 0.7 1 0.5 0.9
-0.06 0.6 0.5 0.4 0.4 0.4 0.5 1 0.6
0.04 0.8 0.8 0.7 0.7 0.7 0.9 0.6 1
1.0
0.8
0.6
0.4
0.2
0.0

Если основной целью работы будет узнать вероятность поступления в магистратуру в европейские ВУЗы, то судя по корреляционной матрице, построение такой модели машинного обучения возможно, и будет достаточно успешным. В целом, можно вкладывать практически все имеющиеся признаки. По корреляционной матрице можно понять, что при построении моделей машинного обучения следует использовать признаки “GRE Score”, “TOEFL Score”, “CGPA”, “Chance of Admit”. Также целевой признак отчасти коррелирует с признаками “University Rating”, “SOP”, “LOR” и “Research”, которые мы также добавим для обучения модели.

Из вышеперечисленных признаков, которые будут включены в модель, наиболее весомый вклад окажут “GRE Score”, “TOEFL Score”, “CGPA”, “Chance of Admit”.