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 «Системы обработки информации и управления»

Рубежный контроль № 1

по курсу “Технологии машинного обучения”

по теме “Технологии разведочного анализа и обработки данных”

Вариант 7

Выполнил:

студент группы ИУ5-63Б

Волгина А. Д.

22.04.21

Проверил:

Гапанюк Ю.Е.

2022 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.

Дополнительное требование: Для произвольной колонки данных построить график "Ящик с усами (boxplot)".

Набор данных:

https://www.kaggle.com/mohansacharya/graduate-admissions (файл Admission_Predict_Ver1.1.csv)

Импортируем модули для работы и загрузим набор данных в переменную data:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

```
[2]  data = pd.read_csv('Admission_Predict_Ver1.1.csv', sep=",")
```

Убедимся, что в нём нет пропусков:

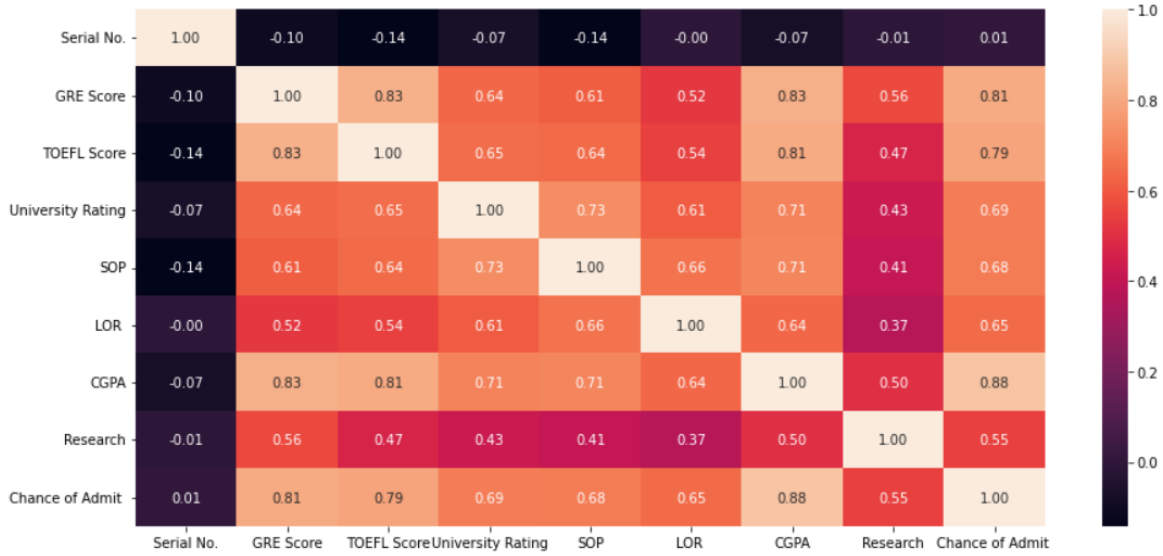
```
[3]  data.isnull().sum()
```

```
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
```

Построим корреляционную матрицу:

```
fig, ax = plt.subplots(figsize=(15,7))
sns.heatmap(data.corr(method='pearson'), ax=ax, annot=True, fmt='.2f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e7b575f50>
```



Признак Serial No стоит выкинуть, потому что от него ничего не зависит. По матрице видно, что параметр “Chance of admit” сильно коррелирует с признаками “GRE Score”, “TOEFL Score” и “CGPA”, с остальными тоже коэффициент корреляции, как минимум, 0,55 - это значит, что мы можем строить линейные модели. Признаки “GRE Score”, “TOEFL Score” и “CGPA” внесут значительный вклад, остальные тоже периодически будут влиять на ключевое значение (“Chance of admit”).

Также выполним дополнительное требование:

```
sns.boxplot(x=data["TOEFL Score"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e78412050>
```