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 №2

Выполнил:

студент группы ИУ5-62Б

Балабанов А.О.

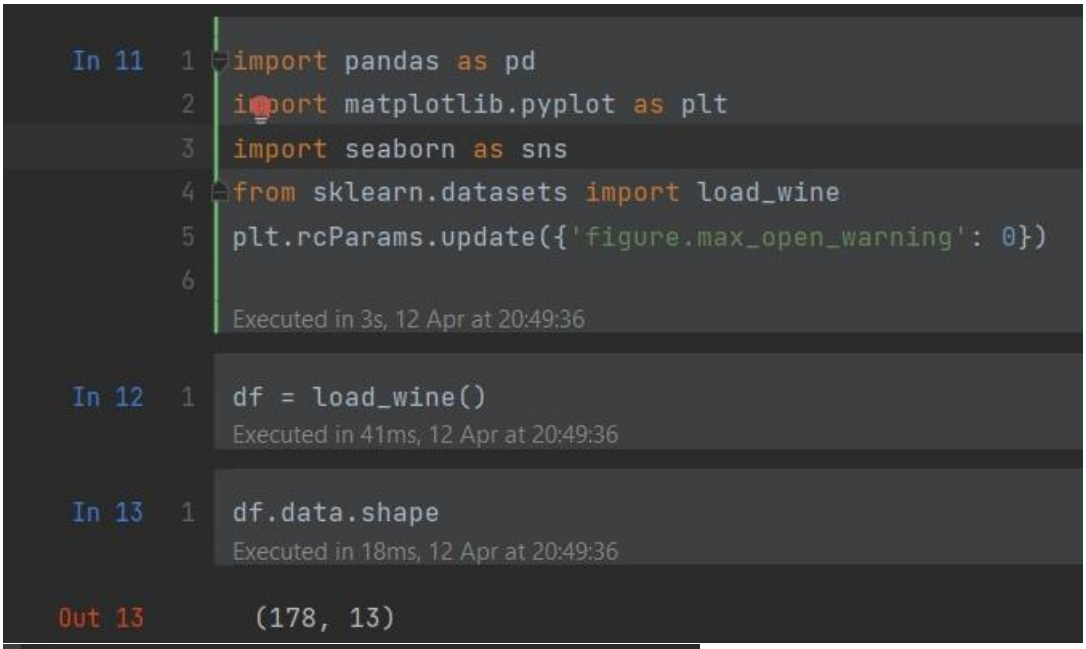
Преподаватель:

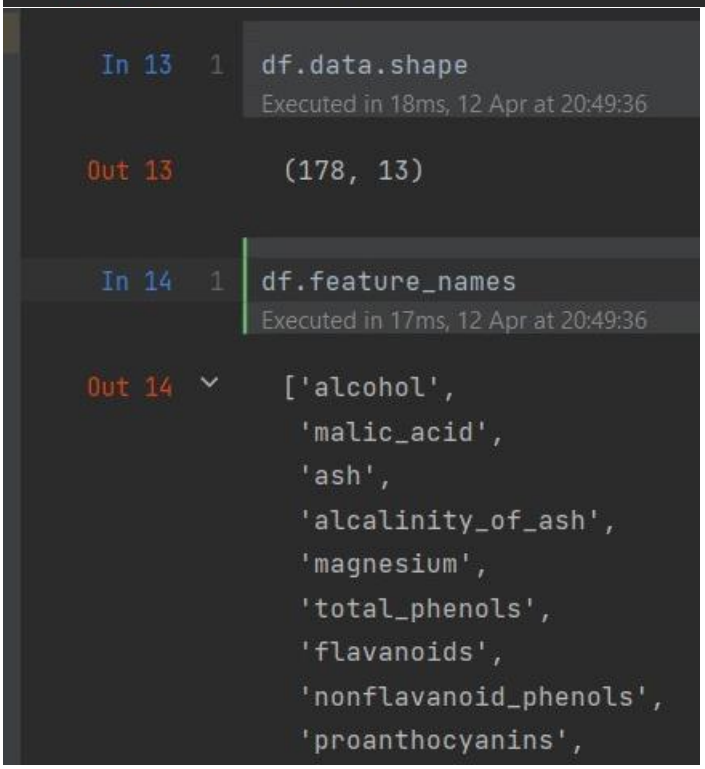
Гапанюк Ю. Е.

2023 г.

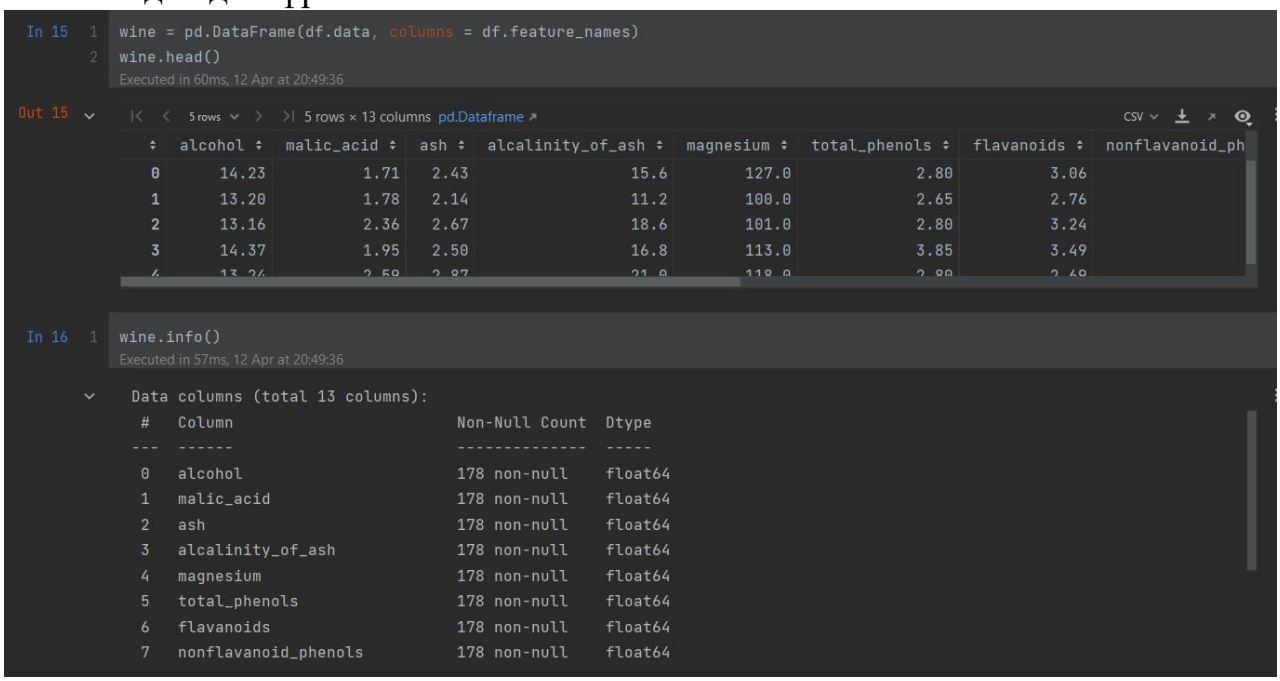
Выполнение работы

Для выполнения задачи проведения корреляционного анализа данных был представлен набор данных sklearn wine





Был создан датафрейм



```
In 15
wine = pd.DataFrame(df.data, columns = df.feature_names)
wine.head()
Executed in 60ms, 12 Apr at 20:49:36
```

Out 15

5 rows × 13 columns pd.Dataframe

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	

```
In 16
wine.info()
Executed in 57ms, 12 Apr at 20:49:36
```

```
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   alcohol                       178 non-null    float64
 1   malic_acid                    178 non-null    float64
 2   ash                           178 non-null    float64
 3   alcalinity_of_ash             178 non-null    float64
 4   magnesium                     178 non-null    float64
 5   total_phenols                 178 non-null    float64
 6   flavanoids                    178 non-null    float64
 7   nonflavanoid_phenols          178 non-null    float64
```

В наборе данных отсутствуют пропуски

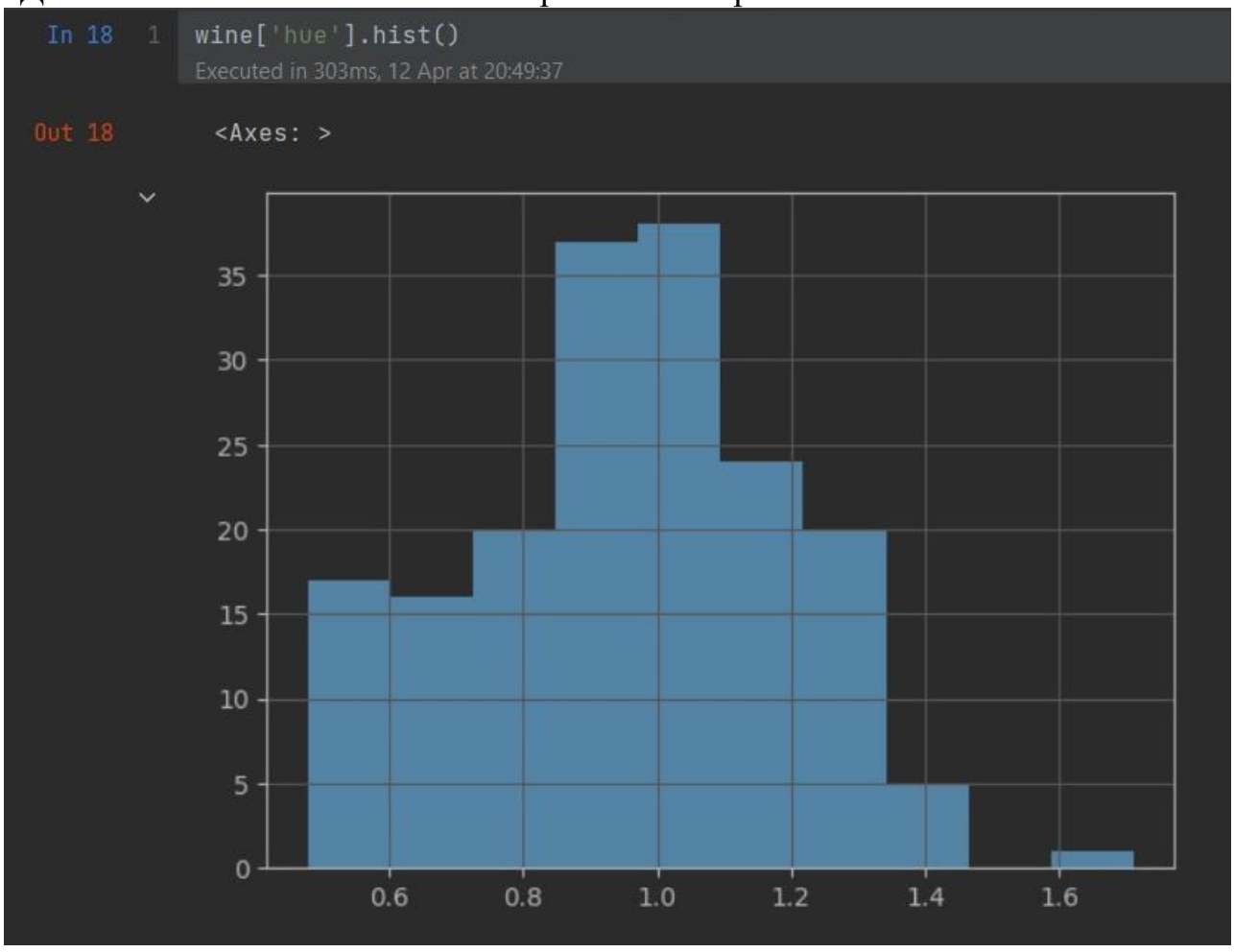
```
In 17  1  wine.isna().sum()
          Executed in 23ms, 12 Apr at 20:49:36
```

Out 17 — 1-10 — Length: 13, dtype: int64 pd.Series — CSV

	<unnamed>
alcohol	0
malic_acid	0
ash	0
alcalinity_of_ash	0
magnesium	0
total_phenols	0
flavanoids	0
nonflavanoid_phenols	0
proanthocyanins	0

```
In 18  1  wine['hue'].hist()
          Executed in 303ms, 12 Apr at 20:49:37
```

Для колонки «оттенки» была построена гистограмма



Для визуализации корреляционной матрицы была использована «тепловая карта»

С целевым признаком наиболее сильную корреляцию имеют признаки “flavanoids” (-0,85), “od280/od315_of_diluted_wines” (-0,79), “total_phenols” (-0,72), “proline” (-0,63) и “hue” (-0,62). Эти признаки будут наиболее информативными при построении моделей машинного обучения. Целевой признак отчасти коррелирует с признаками “alcalinity_of_ash” (0,52), “proanthocyanins” (-0,5), “nonflavanoid_fenols” (0,49) и “malic_acid” (0,44). Эти признаки также стоит использовать при обучении модели. Признаки “alcohol” (- 0,33), “color_intensity” (0,27), “magnesium” (-0,21) и “ash” (-0,05) слабо коррелируют с целевым признаком и могут негативно сказаться на модели машинного обучения, поэтому, скорее всего, их стоит исключить из модели. Но не все признаки, которые имеют сильную и среднюю корреляцию с целевым признаком, стоит использовать для построения модели машинного обучения. Между признаками “flavanoids” и “total_phenols” наблюдается очень сильная корреляция (0,86). Это связано с тем, что флавоноиды относятся к классу полифенолов. Поэтому из этих двух признаков стоит оставить тот, который имеет наибольшую корреляцию с целевым признаком, т.е. “flavanoids”. Остальные нецелевые признаки не коррелируют друг с другом так сильно и между ними не наблюдается почти линейной зависимости. Таким образом, на основе признаков “flavanoids”, “od280/od315_of_diluted_wines”, “proline”, “hue”, “alcalinity_of_ash”, “proanthocyanins”, “nonflavanoid_phenols” и “malic_acid” могут быть построены модели машинного обучения, первые четыре признака могут иметь наиболее весомый вклад в их обучение.