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 ИУ5-63Б

Ахтамбаев Л.Н.

Преподаватель:

Гапанюк Ю. Е.

2023 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 машинного обучения и о возможном вкладе признаков в модель.

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

Набор данных:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris

Решение:

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

Ввод [27]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_wine
plt.rcParams.update({'figure.max_open_warning': 0})
```

Ввод [28]:
```
data = load_wine()
```

Ввод [29]:
```
data.feature_names
```

Out[29]:
```
['alcohol',
 'malic_acid',
 'ash',
 'alcalinity_of_ash',
 'magnesium',
 'total_phenols',
 'flavanoids',
 'nonflavanoid_phenols',
 'proanthocyanins',
 'color_intensity',
 'hue',
 'od280/od315_of_diluted_wines',
 'proline']
```

Создаем датафрейм:

Ввод [30]:
```
data = pd.DataFrame(data=data['data'],columns=data['feature_names'])
```

Ввод [31]:
```
data.head()
```

Out[31]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	

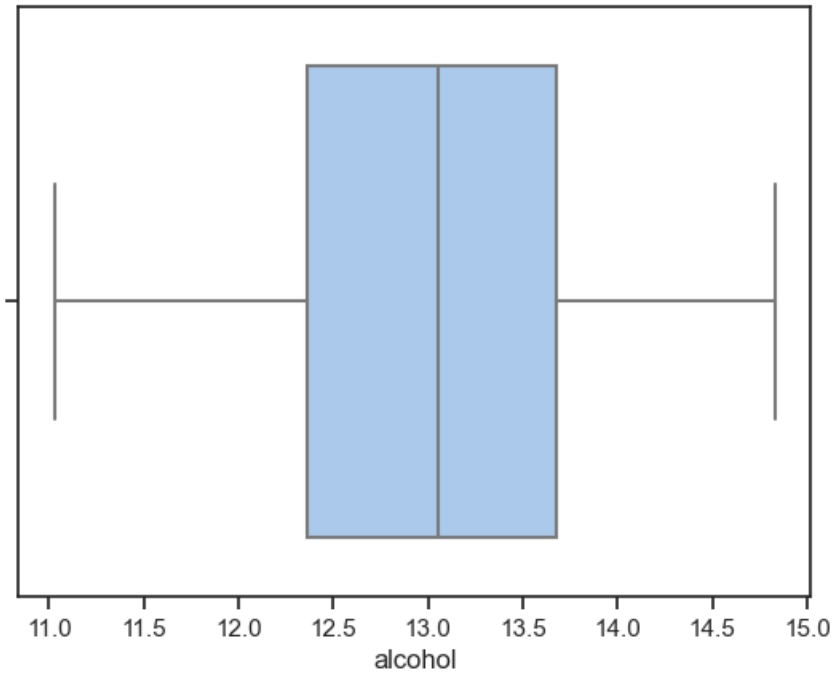
Проверяем набор данных на наличие пропусков:

Пропуски отсутствуют

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

```
Ввод [35]: sns.set_theme(style="ticks", palette="pastel")
           sns.boxplot(x=data["alcohol"])

Out[35]: <Axes: xlabel='alcohol'>
```

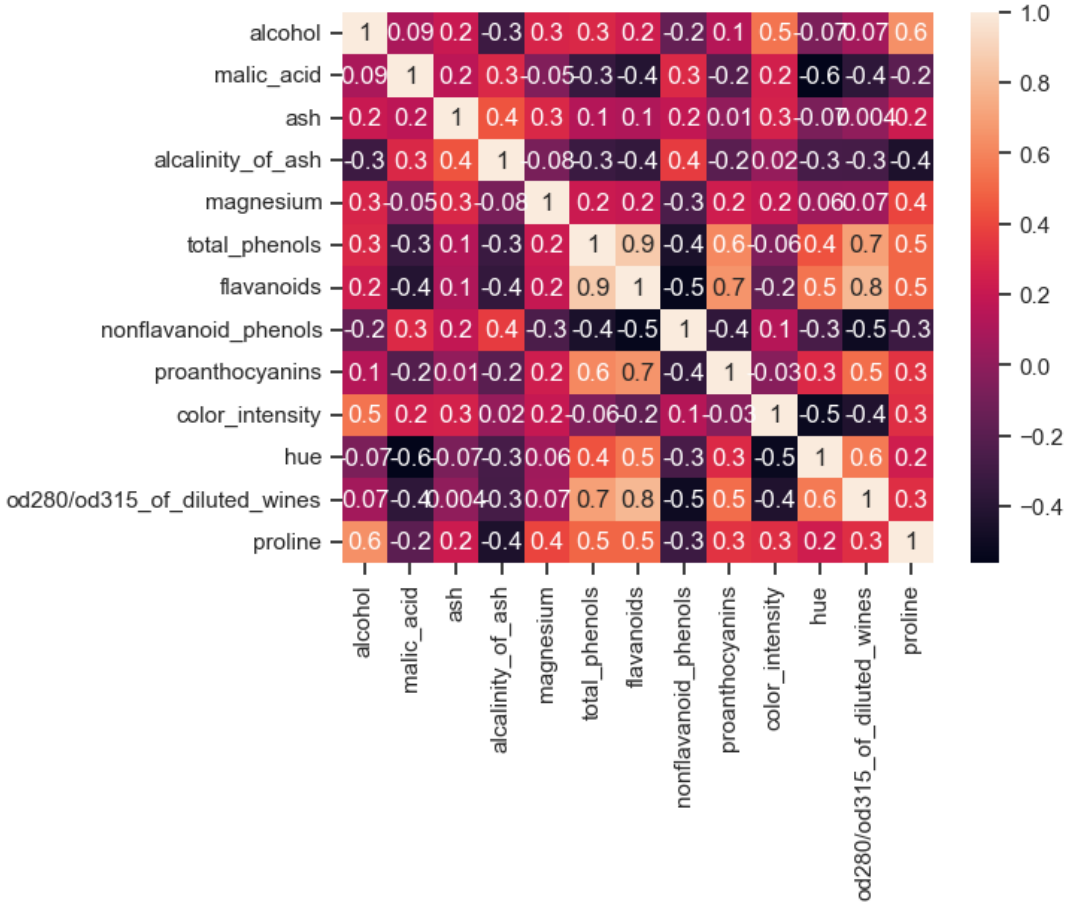


Данный график позволяет увидеть минимальные/максимальные значения, медиану, верхний и нижний квартили.

Перейдем к корреляционному анализу:

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

```
Ввод [40]: sns.heatmap(data.corr(), annot = True, fmt='.1g')
Out[40]: <Axes: >
```



По корреляционной матрице можно понять, что при построении моделей машинного обучения следует использовать признаки “flavanoids” (0.9), “od280/od315_of_diluted_wines” (0.8), “total_phenols” (0.7), “proline” (0.6) и “hue” 0.6), которые имеют наиболее сильную корреляцию с целевым признаком. Также целевой признак отчасти коррелирует с признаками “alcalinity_of_ash” (0.44), “proanthocyanins” (0.5), “nonflavanoid_fenols” (0.5) и “malic_acid” (0.44), которые мы также добавим для обучения модели. Но некоторые признаки, такие как “alcohol” (0.3), “color_intensity” (0.3), “magnesium” (0.2) и “ash” (0.05) слабо коррелируют с целевым признаком и могут негативно сказаться на модели машинного обучения, поэтому их мы включать не будем.

Из двух сильно коррелирующих между собой признаков “flavanoids” и “total_phenols” оставим только “flavanoids”, который имеет наибольшую корреляцию с целевым признаком.

Из вышеперечисленных признаков, которые будут включены в модель, наиболее весомый вклад окажут “flavanoids”, “od280/od315_of_diluted_wines”, “proline” и “hue”.