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

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Рубежный контроль №1

Выполнил:
студент группы ИУ5-62Б
Брусникина Мария
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю.Е.
Подпись и дата:

Москва, 2020 г.

Рубежный контроль №1

Работа Брусникиной М.И., группа ИУ5-62Б, вариант 3 (задача №1, набор данных №3)

Задача

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.

Требование для студентов группы ИУ5-62Б - для произвольной колонки данных построить гистограмму.

Загрузка данных

In [9]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
from sklearn.datasets import *
```

In [10]:

```
wine = load_wine()
for x in wine:
    print(x)
```

```
data
target
target_names
DESCR
feature_names
```

In [11]:

```
data = pd.DataFrame(data= np.c_[wine['data'], wine['target']],
                    columns= wine['feature_names'] + ['target'])
```

In [12]:

```
data.head()
```

Out[12]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_int
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	

In [13]:

```
data.isnull().sum()
```

Out[13]:

```
alcohol                         0
malic_acid                      0
ash                             0
alcalinity_of_ash               0
magnesium                       0
total_phenols                   0
flavanoids                      0
nonflavanoid_phenols            0
proanthocyanins                 0
color_intensity                 0
hue                             0
od280/od315_of_diluted_wines    0
proline                         0
target                          0
dtype: int64
```

Как мы видим, в датасете нет пропусков.

Корреляционный анализ

In [14]:

```
data.corr()
```

Out[14]:

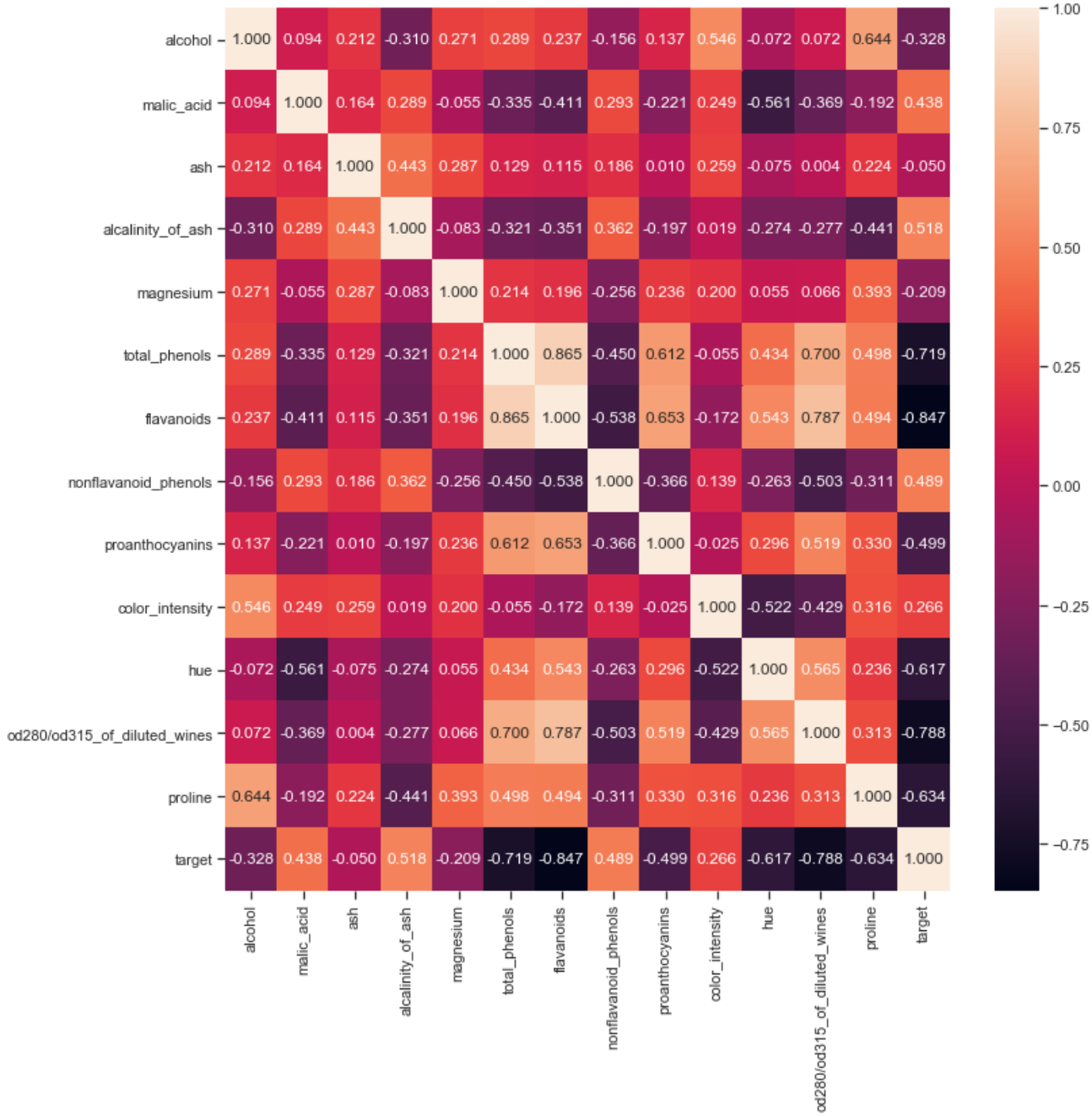
	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.1
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.2
ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.1
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.3
magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.2
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.4
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.5
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.0
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.3
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.1
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.2
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.5
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.3
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.4

In [17]:

```
fig, ax = plt.subplots(figsize=(12,12))
sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x5a22110>
```



На основе корреляционной матрицы можно сделать следующие выводы:

1. Целевой признак наиболее сильно коррелирует с alcalinity_of_ash и nonflavanoid_phenols. Эти признаки обязательно следует оставить в модели.
2. Целевой признак отчасти коррелирует с malic_acid. Этот признак стоит также оставить в модели.
3. Целевой признак слабо коррелирует с color_intensity. Скорее всего этот признак стоит исключить из модели, возможно он только ухудшит качество модели.
4. flavanoids и total_phenols очень сильно коррелируют между собой (0.865). Поэтому из этих признаков в модели можно оставлять только один.

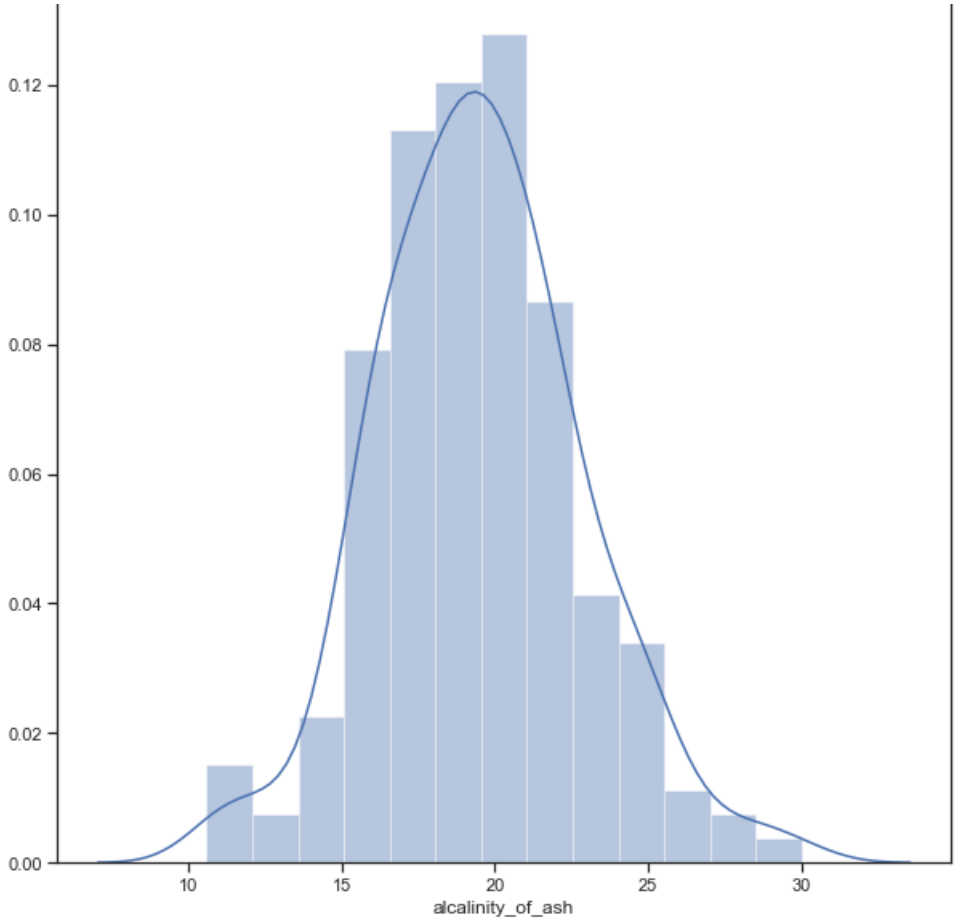
Гистограмма

In [18]:

```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['alcalinity_of_ash'])
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x13145e90>
```



In [ ]: