

Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Дисциплина «Технологии машинного обучения»

Отчёт

по рубежному контролю №1

Тема: «Технологии разведочного анализа и обработки данных.»

Вариант 3

Студент:

Белкина Е.В.

Группа ИУ5-61Б

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 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.

Набор данных:

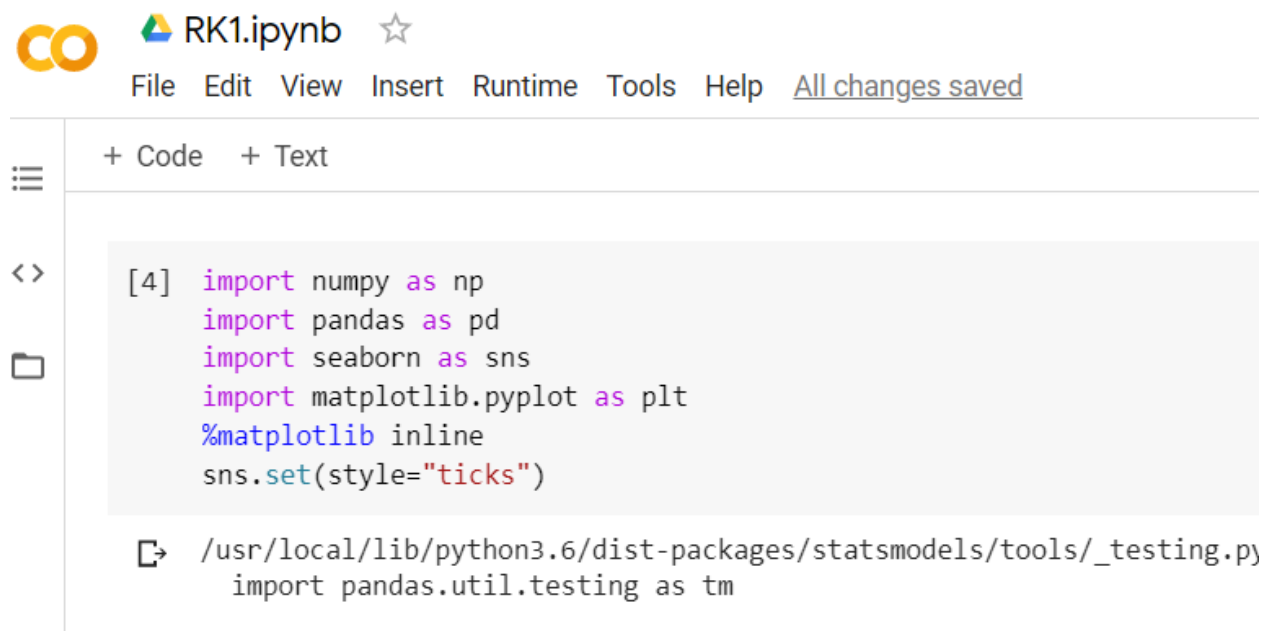
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

Дополнительные требования по группам:

Для студентов групп ИУ5-61Б, ИУ5Ц-81Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Выполнение задания

1. Импортируем необходимые библиотеки с помощью команды `import`.



The screenshot shows a Jupyter Notebook titled "RK1.ipynb". The interface includes a top bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", "Help", and "All changes saved". Below the top bar, there are tabs for "+ Code" and "+ Text". The main area displays a code cell with the following Python code:

```
[4] import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Below the code cell, there is a file explorer showing a file named `/usr/local/lib/python3.6/dist-packages/statsmodels/tools/_testing.py` with the following code snippet:

```
import pandas.util.testing as tm
```

2. Импортируем датасет `load_wine` из `sklearn` в соответствии с заданием варианта

```
[5] from sklearn.datasets import load_wine
wine = load_wine()
```

3. Преобразуем датасет Scikit-learn в Pandas Dataframe

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

4. Проверим наличие пропусков данных

```
# проверим есть ли пропущенные значения
data.isnull().sum()
```

```
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
```

Можем видеть, что пропуски данных в датасете отсутствуют.

Таким образом, мы можем построить корректную корреляционную матрицу.

5. Проведём корреляционный анализ

Матрица с коэффициентом корреляции Пирсона:

data.corr(method='pearson')											
	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.136698	0.546364	-0.071747
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0.220746	0.248985	-0.561296
ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.009652	0.258887	-0.074667
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0.197327	0.018732	-0.273955
magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0.236441	0.199950	0.055398
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.612413	-0.055136	0.433681
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.652692	-0.172379	0.543479
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0.365845	0.139057	-0.262640
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.365845	1.000000	-0.025250	0.295544
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.139057	-0.025250	1.000000	-0.521813
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.262640	0.295544	-0.521813	1.000000
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.503270	0.519067	-0.428815	0.565468
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.311385	0.330417	0.316100	0.236183
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0.499130	0.265668	-0.617369

Матрица с коэффициентом корреляции Кендалла:

data.corr(method='kendall')

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od2
alcohol	1.000000	0.093844	0.170154	-0.212978	0.250506	0.209099	0.191087	-0.109554	0.133526	0.434353	-0.021717	
malic_acid	0.093844	1.000000	0.158178	0.210119	0.050869	-0.174929	-0.211918	0.175129	-0.168714	0.195607	-0.388707	
ash	0.170154	0.158178	1.000000	0.258352	0.254246	0.089855	0.049474	0.098937	0.018240	0.187786	-0.037234	
alcalinity_of_ash	-0.212978	0.210119	0.258352	1.000000	-0.121005	-0.256669	-0.309865	0.278091	-0.171404	-0.057281	-0.239210	
magnesium	0.250506	0.050869	0.254246	-0.121005	1.000000	0.172195	0.161603	-0.158361	0.117871	0.241781	0.023760	
total_phenols	0.209099	-0.174929	0.089855	-0.256669	0.172195	1.000000	0.701999	-0.310443	0.466517	0.028264	0.289210	
flavanoids	0.191087	-0.211918	0.049474	-0.309865	0.161603	0.701999	1.000000	-0.378099	0.534615	0.028674	0.354372	
nonflavanoid_phenols	-0.109554	0.175129	0.098937	0.278091	-0.158361	-0.310443	-0.378099	1.000000	-0.269189	0.036065	-0.179755	
proanthocyanins	0.133526	-0.168714	0.018240	-0.171404	0.117871	0.466517	0.534615	-0.269189	1.000000	-0.014962	0.231071	
color_intensity	0.434353	0.195607	0.187786	-0.057281	0.241781	0.028264	0.028674	0.036065	-0.014962	1.000000	-0.291561	
hue	-0.021717	-0.388707	-0.037234	-0.239210	0.023760	0.289210	0.354372	-0.179755	0.231071	-0.291561	1.000000	
od280/od315_of_diluted_wines	0.061513	-0.162909	-0.006341	-0.226253	0.034307	0.478267	0.520448	-0.363787	0.369104	-0.206046	0.324678	
proline	0.449387	-0.044660	0.171574	-0.313218	0.343016	0.280203	0.263661	-0.174108	0.204172	0.316632	0.143508	
target	-0.238984	0.247494	-0.038085	0.449402	-0.184992	-0.590404	-0.725255	0.379234	-0.450225	0.065124	-0.479229	

Матрица с коэффициентом корреляции Спирмана:

data.corr(method='spearman')

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od2
alcohol	1.000000	0.140430	0.243722	-0.306598	0.365503	0.310920	0.294740	-0.162207	0.192734	0.635425	-0.024203	
malic_acid	0.140430	1.000000	0.230674	0.304069	0.080188	-0.280225	-0.325202	0.255236	-0.244825	0.290307	-0.560265	
ash	0.243722	0.230674	1.000000	0.366374	0.361488	0.132193	0.078796	0.145583	0.024384	0.283047	-0.050183	
alcalinity_of_ash	-0.306598	0.304069	0.366374	1.000000	-0.169558	-0.376657	-0.443770	0.389390	-0.253695	-0.073776	-0.352507	
magnesium	0.365503	0.080188	0.361488	-0.169558	1.000000	0.246417	0.233167	-0.236786	0.173647	0.357029	0.036095	
total_phenols	0.310920	-0.280225	0.132193	-0.376657	0.246417	1.000000	0.879404	-0.448013	0.666689	0.011162	0.439457	
flavanoids	0.294740	-0.325202	0.078796	-0.443770	0.233167	0.879404	1.000000	-0.543897	0.730322	-0.042910	0.535430	
nonflavanoid_phenols	-0.162207	0.255236	0.145583	0.389390	-0.236786	-0.448013	-0.543897	1.000000	-0.384629	0.059639	-0.267813	
proanthocyanins	0.192734	-0.244825	0.024384	-0.253695	0.173647	0.666689	0.730322	-0.384629	1.000000	-0.030947	0.342795	
color_intensity	0.635425	0.290307	0.283047	-0.073776	0.357029	0.011162	-0.042910	0.059639	-0.030947	1.000000	-0.418522	
hue	-0.024203	-0.560265	-0.050183	-0.352507	0.036095	0.439457	0.535430	-0.267813	0.342795	-0.418522	1.000000	
od280/od315_of_diluted_wines	0.103050	-0.255185	-0.007500	-0.325890	0.056963	0.687207	0.741533	-0.494950	0.554031	-0.317516	0.485454	
proline	0.633580	-0.057466	0.253163	-0.456090	0.507575	0.419470	0.429904	-0.270112	0.308249	0.457096	0.207740	
target	-0.354167	0.346913	-0.053988	0.569792	-0.250498	-0.726544	-0.854908	0.474205	-0.570648	0.131170	-0.616570	

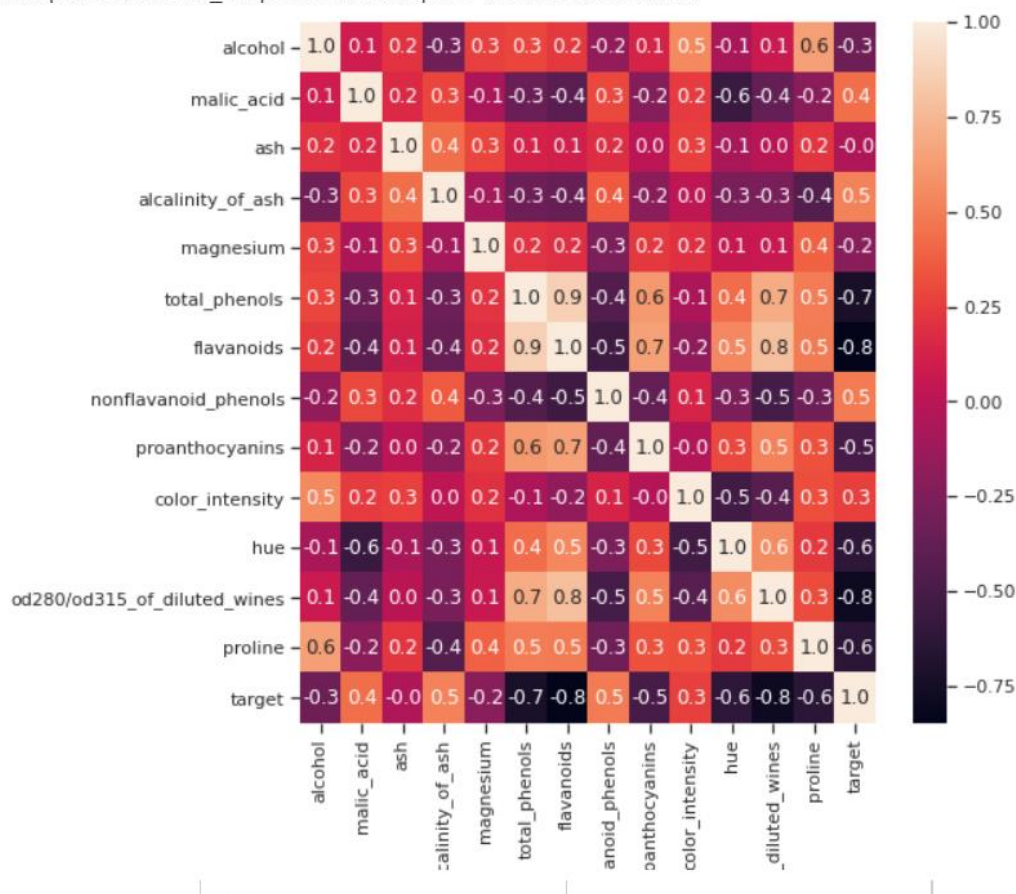
data.corr(method='spearman')

_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
140430	0.243722	-0.306598	0.365503	0.310920	0.294740	-0.162207	0.192734	0.635425	-0.024203	0.103050	0.633580	-0.354167
000000	0.230674	0.304069	0.080188	-0.280225	-0.325202	0.255236	-0.244825	0.290307	-0.560265	-0.255185	-0.057466	0.346913
230674	1.000000	0.366374	0.361488	0.132193	0.078796	0.145583	0.024384	0.283047	-0.050183	-0.007500	0.253163	-0.053988
304069	0.366374	1.000000	-0.169558	-0.376657	-0.443770	0.389390	-0.253695	-0.073776	-0.352507	-0.325890	-0.456090	0.569792
080188	0.361488	-0.169558	1.000000	0.246417	0.233167	-0.236786	0.173647	0.357029	0.036095	0.056963	0.507575	-0.250498
280225	0.132193	-0.376657	0.246417	1.000000	0.879404	-0.448013	0.666689	0.011162	0.439457	0.687207	0.419470	-0.726544
325202	0.078796	-0.443770	0.233167	0.879404	1.000000	-0.543897	0.730322	-0.042910	0.535430	0.741533	0.429904	-0.854908
255236	0.145583	0.389390	-0.236786	-0.448013	-0.543897	1.000000	-0.384629	0.059639	-0.267813	-0.494950	-0.270112	0.474205
244825	0.024384	-0.253695	0.173647	0.666689	0.730322	-0.384629	1.000000	-0.030947	0.342795	0.554031	0.308249	-0.570648
290307	0.283047	-0.073776	0.357029	0.011162	-0.042910	0.059639	-0.030947	1.000000	-0.418522	-0.317516	0.457096	0.131170
560265	-0.050183	-0.352507	0.036095	0.439457	0.535430	-0.267813	0.342795	-0.418522	1.000000	0.485454	0.207740	-0.616570
255185	-0.007500	-0.325890	0.056963	0.687207	0.741533	-0.494950	0.554031	-0.317516	0.485454	1.000000	0.253266	-0.743787
057466	0.253163	-0.456090	0.507575	0.419470	0.429904	-0.270112	0.308249	0.457096	0.207740	0.253266	1.000000	-0.576383
346913	-0.053988	0.569792	-0.250498	-0.726544	-0.854908	0.474205	-0.570648	0.131170	-0.616570	-0.743787	-0.576383	1.000000

Тепловая карта корреляционной матрицы:

```
fig, ax = plt.subplots(figsize=(8,8))
[20] sns.heatmap(data.corr(), annot=True, fmt='.1f')
```

<matplotlib.axes._subplots.AxesSubplot at 0x7ff8e0b8aac8>



Корреляционный анализ:

Необходимо понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (колонка "target"). Именно эти признаки будут наиболее информативными для моделей машинного обучения. Значительное большинство признаков с target коррелируют alkalinity_of_ash (0.5), anoid_phenols (0.5), malic_acid (0.4). Эти признаки следует оставить в модели.

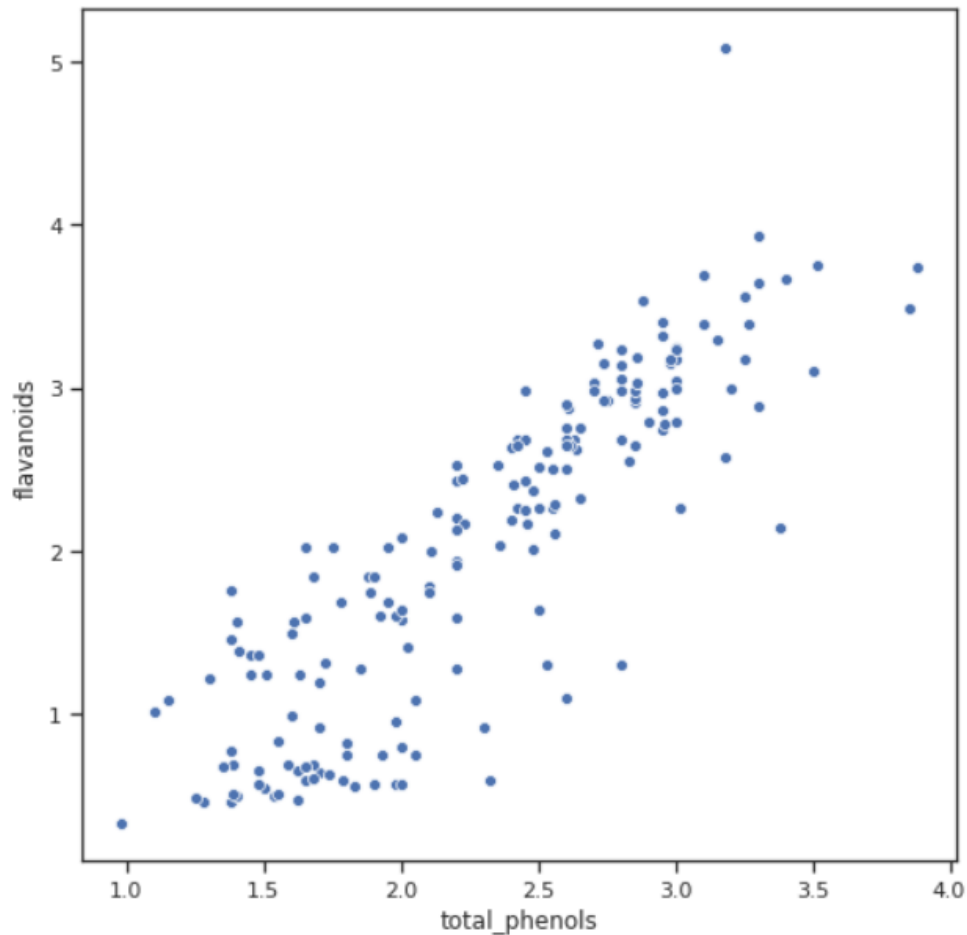
Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. В данном датасете такими являются flavanoids (-0.8), total_phenols (-0.7), od280/od315_of_diluted_wines (-0.8), hue (-0.6), proline (-0.6).

Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если признаки линейно зависимы, то для построения модели из них выбирают какой-то один признак. В нашем наборе данных максимально коррелируют между собой flavonoids и total_phenols (0.9), они имеют практически линейную зависимость. Поэтому для построения модели лучше оставить только один из этих признаков, наиболее коррелирующий с целевым. Но оба эти признака так плохо коррелируют с target, что имеет смысл убрать каждый из них.

Диаграмма рассеяния для колонок `total_phenols` и `flavonoids`:

```
fig, ax = plt.subplots(figsize=(8,8))
sns.scatterplot(ax=ax, x='total_phenols', y='flavonoids', data=data)
```

↳ `<matplotlib.axes._subplots.AxesSubplot at 0x7ff8e06f3278>`



6. Выводы

На данном наборе данных (датасет Wine recognition dataset из sklearn) возможно удачно построить модель машинного обучения, так как он не имеет пропусков данных, содержит признаки, значительно коррелирующие с целевым и не зависящие линейно. Для улучшения построенной модели имеется возможность отказаться от линейно зависимых и слабо коррелирующих с целевым признаков.