Ролдугин Е.В.

ИУ5-25М

Рубежный контроль №1 по курсу Методов Машинного Обучения

ВАРИАНТ 11

Набор данных содержит список Вин. Используются данные из https://www.kaggle.com/datasets/zynicide/wine-reviews

Загрузка и первичный анализ данных

```
[1]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
```

```
[80]: data = pd.read_csv('winemag-data-130k-v2.csv')
```

```
[92]: data.head(8)
```

[92]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	Not Stated	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks
5	5	Spain	Blackberry and raspberry aromas show a typical...	Ars In Vitro	87	15.0	Northern Spain	Navarra	NaN	Michael Schachner	@wineschach	Tandem 2011 Ars In Vitro Tempranillo-Merlot (N...	Tempranillo-Merlot	Tandem
6	6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	Kerin O'Keefe	@kerinokeefe	Terre di Giurfo 2013 Belsito Frappato (Vittoria)	Frappato	Terre di Giurfo
7	7	France	This dry and restrained wine offers spice in p...	NaN	87	24.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Trimbach 2012 Gewurztraminer (Alsace)	Gewürztraminer	Trimbach

[82]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129971 entries, 0 to 129970
Data columns (total 14 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Unnamed: 0             129971 non-null  int64
 1   country                129908 non-null  object
 2   description            129971 non-null  object
 3   designation            92506 non-null   object
 4   points                 129971 non-null  int64
 5   price                  120975 non-null  float64
 6   province               129908 non-null  object
 7   region_1               108724 non-null  object
 8   region_2               50511 non-null   object
 9   taster_name            103727 non-null  object
 10  taster_twitter_handle  98758 non-null   object
 11  title                  129971 non-null  object
 12  variety                129970 non-null  object
 13  winery                 129971 non-null  object
dtypes: float64(1), int64(2), object(11)
memory usage: 8.4+ MB
```

[83]:
```
data.describe()
```

[83]:

	Unnamed: 0	points	price
count	129971.000000	129971.000000	120975.000000
mean	64985.000000	88.447138	35.363389
std	37519.540256	3.039730	41.022218
min	0.000000	80.000000	4.000000
25%	32492.500000	86.000000	17.000000
50%	64985.000000	88.000000	25.000000
75%	97477.500000	91.000000	42.000000
max	129970.000000	100.000000	3300.000000

[86]:
```
data[['country', 'points', 'region_1', 'variety']]
```

[86]:

	country	points	region_1	variety
0	Italy	87	Etna	White Blend
1	Portugal	87	NaN	Portuguese Red
2	US	87	Willamette Valley	Pinot Gris
3	US	87	Lake Michigan Shore	Riesling
4	US	87	Willamette Valley	Pinot Noir
...	...	...	...	...
129966	Germany	90	NaN	Riesling
129967	US	90	Oregon	Pinot Noir
129968	France	90	Alsace	Gewürztraminer
129969	France	90	Alsace	Pinot Gris
129970	France	90	Alsace	Gewürztraminer

129971 rows × 4 columns

Задание №11

Для набора данных проведите устранение пропусков для одного (произвольного) категориального признака с использованием метода заполнения отдельной категорией для пропущенных значений.

В качестве произвольного признака выберем колонку "region_1". Затем заменим пропущенные значения категорией "Not Stated"

```
[87]: data['region_1'].fillna('Not Stated', inplace = True)
```

```
[88]: data['region_1'].isna().sum()
```

```
[88]: 0
```

```
[89]: data.head(20)
```

```
[91]: data[data['region_1'] == 'Not Stated'][['country', 'points', 'region_1', 'variety']]
```

[91]:

	country	points	region_1	variety
1	Portugal	87	Not Stated	Portuguese Red
8	Germany	87	Not Stated	Gewürztraminer
15	Germany	87	Not Stated	Riesling
36	Chile	86	Not Stated	Viognier-Chardonnay
44	Chile	86	Not Stated	Merlot
...	...	...	...	...
129956	New Zealand	90	Not Stated	Bordeaux-style Red Blend
129958	New Zealand	90	Not Stated	Bordeaux-style Red Blend
129960	Portugal	90	Not Stated	Pinot Noir
129963	Israel	90	Not Stated	Cabernet Sauvignon
129966	Germany	90	Not Stated	Riesling

21247 rows × 4 columns

Заметим, что все пропущенные значения были успешно заменены на "Not Stated"

Задание №31

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте метод обертывания (wrapper method), прямой алгоритм (sequential forward selection).

[13]:

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

[14]:

```python
data = pd.read_csv("KNNAlgorithmDataset.csv")
```

[15]:

```python
data[['diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'compactness_mean']]
```

[15]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	compactness_mean
0	M	17.99	10.38	122.80	1001.0	0.27760
1	M	20.57	17.77	132.90	1326.0	0.07864
2	M	19.69	21.25	130.00	1203.0	0.15990
3	M	11.42	20.38	77.58	386.1	0.28390
4	M	20.29	14.34	135.10	1297.0	0.13280
...	...	...	...	...	...	...
564	M	21.56	22.39	142.00	1479.0	0.11590
565	M	20.13	28.25	131.20	1261.0	0.10340
566	M	16.60	28.08	108.30	858.1	0.10230
567	M	20.60	29.33	140.10	1265.0	0.27700
568	B	7.76	24.54	47.92	181.0	0.04362

569 rows × 6 columns

[ ]:

```python
from sklearn.neighbors import KNeighborsClassifier
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
```

Выберем “diagnosis” для предсказания прогноза

[ ]:

```python
X = data.drop(labels = 'diagnosis', axis = 1).copy(deep = True)
Y = data['diagnosis'].copy(deep = True)
knn = KNeighborsClassifier(n_neighbors=5)
sfs = SFS(knn, forward = True, floating = False, k_feature = 4)
```

[ ]:

```python
sfs.fit(X,Y)
```

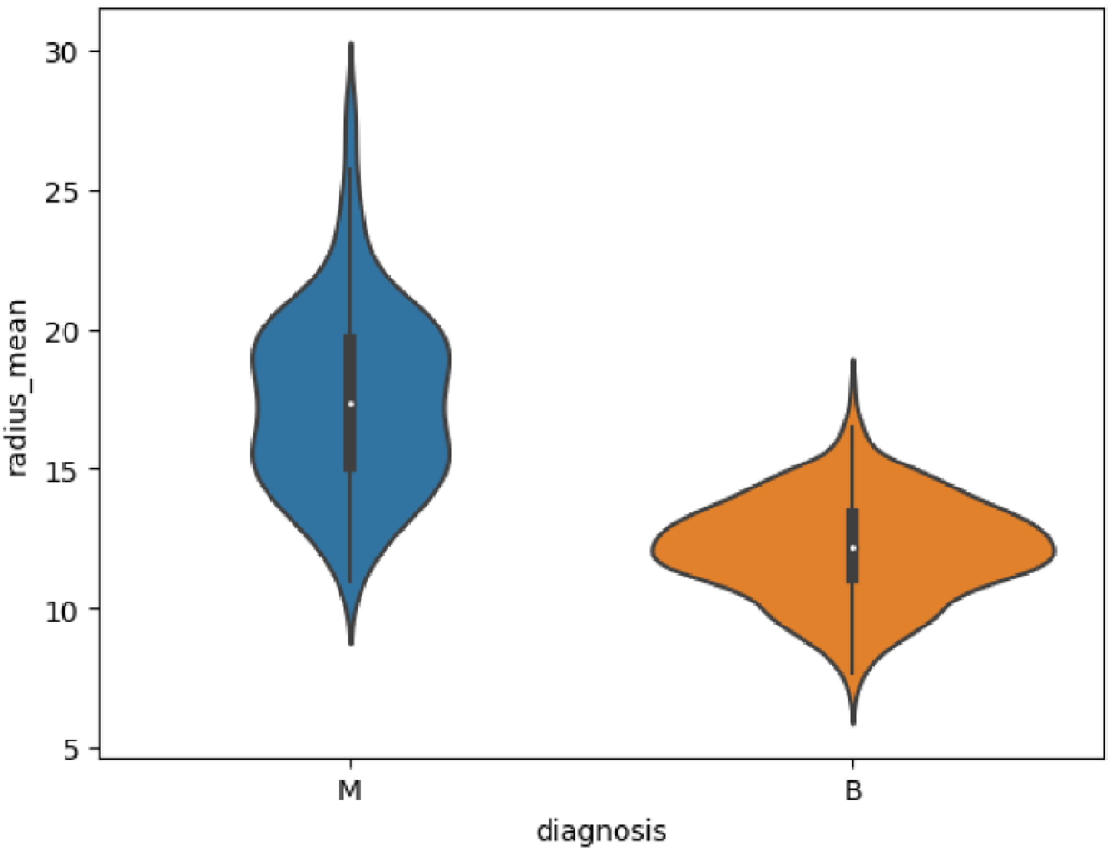
```
▸ SequentialFeatureSelector
▸ estimator: KNeighborsClassifier
▸ KNeighborsClassifier
```

```
[ ]: sfs.subsets_
```

```
{1: {'feature_idx': (7,),
  'cv_scores': array([0.86842105, 0.9122807 , 0.9122807 , 0.92982456, 0.90265487]),
  'avg_score': 0.9050923769600994,
  'feature_names': ('concave points_mean',)},
 2: {'feature_idx': (7, 16),
  'cv_scores': array([0.92105263, 0.93859649, 0.90350877, 0.93859649, 0.90265487]),
  'avg_score': 0.9208818506443098,
  'feature_names': ('concave points_mean', 'concavity_se')},
 3: {'feature_idx': (7, 16, 20),
  'cv_scores': array([0.85087719, 0.92105263, 0.93859649, 0.94736842, 0.9380531 ]),
  'avg_score': 0.9191895668374475,
  'feature_names': ('concave points_mean', 'concavity_se', 'radius_worst')},
 4: {'feature_idx': (7, 16, 20, 26),
  'cv_scores': array([0.92105263, 0.92982456, 0.95614035, 0.93859649, 0.94690265]),
  'avg_score': 0.9385033379909953,
  'feature_names': ('concave points_mean',
   'concavity_se',
   'radius_worst',
   'concavity_worst')}}
```

Наилучшая точность достигается при выборе признаков 'concave_points_mean', 'concavity_se', 'radius_worst', 'concavity_worst'

```
[ ]: sns.violinplot(data = data, x = 'diagnosis', y = 'radius_mean')
```



Задание для группы ИУ5-25М - для произвольной колонки данных построить парные диаграммы (pairplot).

```
[ ]  sns.pairplot(df[['ratings', 'Number of ratings']], height=3, aspect=2)
```

```
<seaborn.axisgrid.PairGrid at 0x7fdb7e69cd30>
```