Уралова Е.А.

ИУ5-65Б Вариант №16

Импортируем библиотеки:

```
import numpy as np
import pandas as pd
from typing import Dict, Tuple
from scipy import stats
from IPython.display import Image
from sklearn.datasets import load_iris, load_boston
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, balanced_accuracy_score
from sklearn.metrics import precision_score, recall_score, f1_score,
classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import mean_absolute_error, mean_squared_error,
mean_squared_log_error, median_absolute_error, r2_score
from sklearn.metrics import roc_curve, roc_auc_score
from sklearn.preprocessing import MinMaxScaler
from sklearn.datasets import make_blobs, make_circles
from sklearn.model_selection import cross_val_score, cross_validate
from sklearn.svm import SVC, NuSVC, LinearSVC, OneClassSVM, SVR,
NuSVR, LinearSVR
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import AdaBoostClassifier
from sklearn import svm
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
data = pd.read_csv('restaurant-scores-lives-standard.csv', sep=",")
```

```
data.head()
```

```
   business_id          business_name          business_address  \
0       101192           Cochinita #2  2 Marina Blvd Fort Mason
1        97975             BREADBELLY           1408 Clement St
2        92982  Great Gold Restaurant             3161 24th St.
3       101389                 HOMAGE          214 CALIFORNIA ST
4        85986           Pronto Pizza               798 Eddy St

   business_city business_state business_postal_code
business_latitude  \
0  San Francisco             CA                  NaN
NaN
```

```
1  San Francisco           CA              94118
NaN
2  San Francisco           CA              94110
NaN
3  San Francisco           CA              94111
NaN
4  San Francisco           CA              94109
NaN

   business_longitude business_location  business_phone_number  ...  \
0                 NaN               NaN           1.415043e+10  ...
1                 NaN               NaN           1.415724e+10  ...
2                 NaN               NaN                    NaN  ...
3                 NaN               NaN           1.415488e+10  ...
4                 NaN               NaN                    NaN  ...

         inspection_type           violation_id  \
0          New Ownership                    NaN
1  Routine - Unscheduled  97975_20190725_103124
2          New Ownership                    NaN
3       New Construction                    NaN
4          New Ownership  85986_20161011_103114

                              violation_description  risk_category  \
0                                               NaN            NaN
1  Inadequately cleaned or sanitized food contact...  Moderate Risk
2                                               NaN            NaN
3                                               NaN            NaN
4                     High risk vermin infestation      High Risk

  Neighborhoods (old) Police Districts Supervisor Districts  \
0                 NaN              NaN                  NaN
1                 NaN              NaN                  NaN
2                 NaN              NaN                  NaN
3                 NaN              NaN                  NaN
4                 NaN              NaN                  NaN

   Fire Prevention Districts  Zip Codes  Analysis Neighborhoods
0                        NaN        NaN                     NaN
1                        NaN        NaN                     NaN
2                        NaN        NaN                     NaN
3                        NaN        NaN                     NaN
4                        NaN        NaN                     NaN

[5 rows x 23 columns]
```

Обработка пропусков

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

```
business_id                       0
business_name                     0
business_address                  0
business_city                     0
business_state                    0
business_postal_code           1018
business_latitude             19556
business_longitude            19556
business_location             19556
business_phone_number         36938
inspection_id                     0
inspection_date                   0
inspection_score              13610
inspection_type                   0
violation_id                  12870
violation_description         12870
risk_category                 12870
Neighborhoods (old)           19594
Police Districts              19594
Supervisor Districts          19594
Fire Prevention Districts     19646
Zip Codes                     19576
Analysis Neighborhoods        19594
dtype: int64
```

```
data.shape
```

```
(53973, 23)
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 53973
```

```
data = data.dropna(axis=0, how='any')
data.shape
```

```
(6566, 23)
```

```
data.head()
```

```
     business_id                business_name          business_address
\
11          4794                     VICTOR'S          210 TOWNSEND St

172        63652  SFDH - Banquet Main Kitchen  450 Powell St 2nd Floor

327          328                       Miyako          1470 Fillmore St

372         2684            ERIC'S RESTAURANT            1500 Church St
```

```
397          328                       Miyako        1470 Fillmore St


     business_city business_state business_postal_code
business_latitude  \
11   San Francisco             CA                94107
37.778634
172  San Francisco             CA                94102
37.788918
327  San Francisco             CA                94115
37.783017
372  San Francisco             CA                94131
37.746759
397  San Francisco             CA                94115
37.783017

     business_longitude
business_location  \
11          -122.393089  {'type': 'Point', 'coordinates': [-
122.393089,...
172         -122.408507  {'type': 'Point', 'coordinates': [-
122.408507,...
327         -122.432584  {'type': 'Point', 'coordinates': [-
122.432584,...
372         -122.426995  {'type': 'Point', 'coordinates': [-
122.426995,...
397         -122.432584  {'type': 'Point', 'coordinates': [-
122.432584,...

     business_phone_number  ...        inspection_type
violation_id  \
11            1.415561e+10  ...  Routine - Unscheduled
4794_20181030_103138
172           1.415540e+10  ...  Routine - Unscheduled
63652_20190904_103133
327           1.415554e+10  ...  Routine - Unscheduled
328_20161122_103103
372           1.415528e+10  ...  Routine - Unscheduled
2684_20190715_103109
397           1.415554e+10  ...  Routine - Unscheduled
328_20161122_103149

                                 violation_description  risk_category
\
11   Improper storage use or identification of toxi...       Low Risk

172              Foods not protected from contamination  Moderate Risk

327                  High risk food holding temperature      High Risk
```

```
372        Unclean or unsanitary food contact surfaces      High Risk

397  Wiping cloths not clean or properly stored or ...       Low Risk


    Neighborhoods (old) Police Districts Supervisor Districts  \
11                 34.0              2.0                  9.0
172                 6.0              1.0                 10.0
327                41.0              9.0                 11.0
372                22.0              7.0                  5.0
397                41.0              9.0                 11.0

     Fire Prevention Districts  Zip Codes  Analysis Neighborhoods
11                         6.0    28856.0                    34.0
172                        5.0    28852.0                     8.0
327                       15.0    29490.0                    39.0
372                        2.0       63.0                    22.0
397                       15.0    29490.0                    39.0

[5 rows x 23 columns]
```

Кодируем категориальные признаки

Удалим колонки, которые не влияют на целевой признак:

```
data = data.drop(columns='business_name')
data = data.drop(columns='business_address')
data = data.drop(columns='business_city')
data = data.drop(columns='business_state')
data = data.drop(columns='business_location')
data = data.drop(columns='business_phone_number')
data = data.drop(columns='violation_description')
```

```
data.shape
```

```
(6566, 16)
```

```
data.head()
```

```
     business_id business_postal_code  business_latitude
business_longitude  \
11          4794                94107          37.778634         -
122.393089
172        63652                94102          37.788918         -
122.408507
327          328                94115          37.783017         -
122.432584
372         2684                94131          37.746759         -
122.426995
```

```
397          328                94115          37.783017         -
122.432584

      inspection_id          inspection_date  inspection_score  \
11    4794_20181030  2018-10-30T00:00:00.000              71.0   
172  63652_20190904  2019-09-04T00:00:00.000              94.0   
327    328_20161122  2016-11-22T00:00:00.000              81.0   
372   2684_20190715  2019-07-15T00:00:00.000              87.0   
397    328_20161122  2016-11-22T00:00:00.000              81.0   

           inspection_type           violation_id  risk_category  \
11   Routine - Unscheduled   4794_20181030_103138       Low Risk   
172  Routine - Unscheduled  63652_20190904_103133  Moderate Risk   
327  Routine - Unscheduled    328_20161122_103103      High Risk   
372  Routine - Unscheduled   2684_20190715_103109      High Risk   
397  Routine - Unscheduled    328_20161122_103149       Low Risk   

     Neighborhoods (old)  Police Districts  Supervisor Districts  \
11                  34.0               2.0                   9.0   
172                  6.0               1.0                  10.0   
327                 41.0               9.0                  11.0   
372                 22.0               7.0                   5.0   
397                 41.0               9.0                  11.0   

     Fire Prevention Districts  Zip Codes  Analysis Neighborhoods  
11                         6.0    28856.0                    34.0  
172                        5.0    28852.0                     8.0  
327                       15.0    29490.0                    39.0  
372                        2.0       63.0                    22.0  
397                       15.0    29490.0                    39.0  

data.dtypes

business_id                    int64
business_postal_code          object
business_latitude            float64
business_longitude           float64
inspection_id                 object
inspection_date               object
inspection_score             float64
inspection_type               object
violation_id                  object
risk_category                 object
Neighborhoods (old)          float64
Police Districts             float64
Supervisor Districts         float64
Fire Prevention Districts    float64
Zip Codes                    float64
Analysis Neighborhoods       float64
dtype: object
```

Кодируем категориальные признаки:

```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
le = LabelEncoder()
df_int = le.fit_transform(data['business_postal_code'])
data['business_postal_code'] = df_int
df_int = le.fit_transform(data['inspection_id'])
data['inspection_id'] = df_int
df_int = le.fit_transform(data['inspection_date'])
data['inspection_date'] = df_int
df_int = le.fit_transform(data['inspection_type'])
data['inspection_type'] = df_int
df_int = le.fit_transform(data['violation_id'])
data['violation_id'] = df_int
df_int = le.fit_transform(data['risk_category'])
data['risk_category'] = df_int
data.head()
```

```
     business_id  business_postal_code  business_latitude
business_longitude  \
11          4794                     5          37.778634         -
122.393089
172        63652                     1          37.788918         -
122.408507
327          328                    13          37.783017         -
122.432584
372         2684                    22          37.746759         -
122.426995
397          328                    13          37.783017         -
122.432584

     inspection_id  inspection_date  inspection_score  inspection_type
\
11             829              440              71.0                0

172           1335              604              94.0                0

327            564               28              81.0                0

372            405              576              87.0                0

397            564               28              81.0                0


     violation_id  risk_category  Neighborhoods (old)  Police
Districts  \
11           2734              1                 34.0
2.0
172          3895              2                  6.0
1.0
```

```
327          1925              0                 41.0
9.0
372          1406              0                 22.0
7.0
397          1929              1                 41.0
9.0

     Supervisor Districts  Fire Prevention Districts  Zip Codes  \
11                    9.0                        6.0    28856.0
172                  10.0                        5.0    28852.0
327                  11.0                       15.0    29490.0
372                   5.0                        2.0       63.0
397                  11.0                       15.0    29490.0

     Analysis Neighborhoods
11                     34.0
172                     8.0
327                    39.0
372                    22.0
397                    39.0
```

Масштабируем числовые данные:

```
sc1 = MinMaxScaler()
data['business_id'] = sc1.fit_transform(data[['business_id']])
data['business_latitude'] =
sc1.fit_transform(data[['business_latitude']])
data['business_longitude'] =
sc1.fit_transform(data[['business_longitude']])
data['inspection_score'] =
sc1.fit_transform(data[['inspection_score']])
data['Neighborhoods (old)'] = sc1.fit_transform(data[['Neighborhoods
(old)']])
data['Police Districts'] = sc1.fit_transform(data[['Police
Districts']])
data['Supervisor Districts'] = sc1.fit_transform(data[['Supervisor
Districts']])
data['Fire Prevention Districts'] = sc1.fit_transform(data[['Fire
Prevention Districts']])
data['Zip Codes'] = sc1.fit_transform(data[['Zip Codes']])
data['Analysis Neighborhoods'] = sc1.fit_transform(data[['Analysis
Neighborhoods']])
data.head()
```

```
     business_id  business_postal_code  business_latitude
business_longitude  \
11      0.065966                     5           0.700222
0.903650
172     0.885088                     1           0.803468
0.784522
327     0.003813                    13           0.744225
```

```
0.598490
372     0.036601                        22           0.380214
0.641674
397     0.003813                        13           0.744225
0.598490

     inspection_id  inspection_date  inspection_score  inspection_type
\
11             829              440          0.480769                0

172           1335              604          0.923077                0

327            564               28          0.673077                0

372            405              576          0.788462                0

397            564               28          0.673077                0


     violation_id  risk_category  Neighborhoods (old)  Police
Districts  \
11           2734              1                0.825
0.111111
172          3895              2                0.125
0.000000
327          1925              0                1.000
0.888889
372          1406              0                0.525
0.666667
397          1929              1                1.000
0.888889

     Supervisor Districts  Fire Prevention Districts  Zip Codes  \
11                    0.8                   0.357143   0.978395
172                   0.9                   0.285714   0.978259
327                   1.0                   1.000000   0.999932
372                   0.4                   0.071429   0.000306
397                   1.0                   1.000000   0.999932

     Analysis Neighborhoods
11                    0.825
172                   0.175
327                   0.950
372                   0.525
397                   0.950
```

Делим выборку на обучающую и тестовую

```
target = data['risk_category']
data_X_train, data_X_test, data_y_train, data_y_test =
train_test_split(
    data, target, test_size=0.2, random_state=1)
```

```
data_X_train.shape, data_y_train.shape
```

```
((5252, 16), (5252,))
```

```
data_X_test.shape, data_y_test.shape
```

```
((1314, 16), (1314,))
```

```
np.unique(target)
```

```
array([0, 1, 2])
```

Метод опорных векторов

```
svr_1 = LinearSVC()
svr_1.fit(data_X_train, data_y_train)
```

```
C:\Users\Asus\anaconda3\lib\site-packages\sklearn\svm\_base.py:985:
ConvergenceWarning: Liblinear failed to converge, increase the number
of iterations.
  warnings.warn("Liblinear failed to converge, increase "
```

```
LinearSVC()
```

```
data_y_pred_1 = svr_1.predict(data_X_test)
```

```
accuracy_score(data_y_test, data_y_pred_1)
```

```
0.8599695585996956
```

```
f1_score(data_y_test, data_y_pred_1, average='micro')
```

```
0.8599695585996956
```

```
f1_score(data_y_test, data_y_pred_1, average='macro')
```

```
0.673469111760251
```

```
f1_score(data_y_test, data_y_pred_1, average='weighted')
```

```
0.811584399054054
```

```
svr_2 = LinearSVC(C=1.0, max_iter=10000)
svr_2.fit(data_X_train, data_y_train)
```

```
C:\Users\Asus\anaconda3\lib\site-packages\sklearn\svm\_base.py:985:
ConvergenceWarning: Liblinear failed to converge, increase the number
```

```
of iterations.
  warnings.warn("Liblinear failed to converge, increase "

LinearSVC(max_iter=10000)

data_y_pred_2 = svr_2.predict(data_X_test)

accuracy_score(data_y_test, data_y_pred_2)

0.9984779299847792

f1_score(data_y_test, data_y_pred_2, average='micro')

0.9984779299847792

f1_score(data_y_test, data_y_pred_2, average='macro')

0.9987823618273689

f1_score(data_y_test, data_y_pred_2, average='weighted')

0.998478361107127

svr_3 = LinearSVC(C=1.0, penalty='l1', dual=False, max_iter=10000)
svr_3.fit(data_X_train, data_y_train)

LinearSVC(dual=False, max_iter=10000, penalty='l1')

data_y_pred_3_0 = svr_3.predict(data_X_train)
accuracy_score(data_y_train, data_y_pred_3_0)

0.9996191926884996

data_y_pred_3 = svr_3.predict(data_X_test)
accuracy_score(data_y_test, data_y_pred_3)

1.0

f1_score(data_y_test, data_y_pred_3, average='micro')

1.0

f1_score(data_y_test, data_y_pred_3, average='macro')

1.0

f1_score(data_y_test, data_y_pred_3, average='weighted')

1.0
```

Градиентный бустинг

```
ab1 = AdaBoostClassifier()
ab1.fit(data_X_train, data_y_train)
```

```
AdaBoostClassifier()
```

```
data_y_pred_1 = ab1.predict(data_X_test)
```

```
data_y_pred_1_0 = ab1.predict(data_X_train)
```

```
accuracy_score(data_y_train, data_y_pred_1_0)
```

```
1.0
```

```
accuracy_score(data_y_test, data_y_pred_1)
```

```
1.0
```

```
f1_score(data_y_test, data_y_pred_1, average='micro')
```

```
1.0
```

```
f1_score(data_y_test, data_y_pred_1, average='macro')
```

```
1.0
```

```
f1_score(data_y_test, data_y_pred_1, average='weighted')
```

```
1.0
```

Градиентный бустинг показал лучше качество, чем метод опорных векторов (хотя и у этого метода показатели хорошие)