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**

О Т Ч Е Т
по лабораторной работе № 4

Дисциплина: Технологии машинного обучения
Тема: «Линейные модели, SVM и деревья решений»

Студент	ИУ5Ц-82Б		А.Н. Свинцов
	(Группа)	(Подпись, дата)	(И.О. Фамилия)
Преподаватель			Ю.Е. Гапанюк
		(Подпись, дата)	(И.О. Фамилия)
Преподаватель			А.Н. Нардид
		(Подпись, дата)	(И.О. Фамилия)

Москва, 2023

Лабораторная работа №4

Линейные модели, SVM и деревья решений

Импорт библиотек

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import SGDClassifier
from sklearn.metrics import f1_score, precision_score
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.model_selection import GridSearchCV
import matplotlib.pyplot as plt

target_col='class'

%matplotlib inline
sns.set(style="ticks")
```

Загрузка датасета

```
data = pd.read_csv('mushrooms.csv')

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   class                     8124 non-null   object
 1   cap-shape                 8124 non-null   object
 2   cap-surface               8124 non-null   object
 3   cap-color                 8124 non-null   object
 4   bruises                   8124 non-null   object
 5   odor                      8124 non-null   object
 6   gill-attachment           8124 non-null   object
 7   gill-spacing              8124 non-null   object
 8   gill-size                 8124 non-null   object
 9   gill-color                8124 non-null   object
 10  stalk-shape               8124 non-null   object
 11  stalk-root                8124 non-null   object
 12  stalk-surface-above-ring  8124 non-null   object
 13  stalk-surface-below-ring  8124 non-null   object
 14  stalk-color-above-ring    8124 non-null   object
 15  stalk-color-below-ring    8124 non-null   object
```

```
 16  veil-type                8124 non-null   object
 17  veil-color               8124 non-null   object
 18  ring-number              8124 non-null   object
 19  ring-type                8124 non-null   object
 20  spore-print-color        8124 non-null   object
 21  population               8124 non-null   object
 22  habitat                  8124 non-null   object
dtypes: object(23)
memory usage: 1.4+ MB

data.head()

  class cap-shape cap-surface cap-color bruises odor gill-
attachment  \
0     p         x           s         n       t    p              f

1     e         x           s         y       t    a              f

2     e         b           s         w       t    l              f

3     p         x           y         w       t    p              f

4     e         x           s         g       f    n              f


  gill-spacing gill-size gill-color  ... stalk-surface-below-ring  \
0            c         n          k  ...                         s
1            c         b          k  ...                         s
2            c         b          n  ...                         s
3            c         n          n  ...                         s
4            w         b          k  ...                         s

  stalk-color-above-ring stalk-color-below-ring veil-type veil-
color  \
0                      w                      w         p          w

1                      w                      w         p          w

2                      w                      w         p          w

3                      w                      w         p          w

4                      w                      w         p          w


  ring-number ring-type spore-print-color population habitat
0           o         p                 k          s       u
1           o         p                 n          n       g
2           o         p                 n          n       m
```

```
3           o         p                    k          s       u
4           o         e                    n          a       g

[5 rows x 23 columns]
```

```
data.shape
```

```
(8124, 23)
```

```
#Проверка на пропуски
data.isnull().sum()
```

```
class                       0
cap-shape                   0
cap-surface                 0
cap-color                   0
bruises                     0
odor                        0
gill-attachment             0
gill-spacing                0
gill-size                   0
gill-color                  0
stalk-shape                 0
stalk-root                  0
stalk-surface-above-ring    0
stalk-surface-below-ring    0
stalk-color-above-ring      0
stalk-color-below-ring      0
veil-type                   0
veil-color                  0
ring-number                 0
ring-type                   0
spore-print-color           0
population                  0
habitat                     0
dtype: int64
```

Пустых строк нет

```
Кодируем категориальные признаки
for col in data.columns:
    null_count = data[data[col].isnull()].shape[0]
    if null_count == 0:
        column_type = data[col].dtype
        print('{} - {} - {}'.format(col, column_type, null_count))
```

```
class - object - 0
cap-shape - object - 0
cap-surface - object - 0
cap-color - object - 0
bruises - object - 0
```

```
odor - object - 0
gill-attachment - object - 0
gill-spacing - object - 0
gill-size - object - 0
gill-color - object - 0
stalk-shape - object - 0
stalk-root - object - 0
stalk-surface-above-ring - object - 0
stalk-surface-below-ring - object - 0
stalk-color-above-ring - object - 0
stalk-color-below-ring - object - 0
veil-type - object - 0
veil-color - object - 0
ring-number - object - 0
ring-type - object - 0
spore-print-color - object - 0
population - object - 0
habitat - object - 0
```

```
le = LabelEncoder()
for col in data.columns:
    column_type = data[col].dtype
    if column_type == 'object':
        data[col] = le.fit_transform(data[col]);
        print(col)
```

```
class
cap-shape
cap-surface
cap-color
bruises
odor
gill-attachment
gill-spacing
gill-size
gill-color
stalk-shape
stalk-root
stalk-surface-above-ring
stalk-surface-below-ring
stalk-color-above-ring
stalk-color-below-ring
veil-type
veil-color
ring-number
ring-type
spore-print-color
population
habitat
```

Обучающая и тестовая выборки

```
X = data.drop(target_col, axis=1)
Y = data[target_col]
```

```
X
```

```
      cap-shape  cap-surface  cap-color  bruises  odor  gill-
attachment  \
0             5            2          4        1     6
1
1             5            2          9        1     0
1
2             0            2          8        1     3
1
3             5            3          8        1     6
1
4             5            2          3        0     5
1
...         ...          ...        ...      ...   ...             ..
.
8119          3            2          4        0     5
0
8120          5            2          4        0     5
0
8121          2            2          4        0     5
0
8122          3            3          4        0     8
1
8123          5            2          4        0     5
0

      gill-spacing  gill-size  gill-color  stalk-shape  ...  \
0                0          1           4            0  ...
1                0          0           4            0  ...
2                0          0           5            0  ...
3                0          1           5            0  ...
4                1          0           4            1  ...
...            ...        ...         ...          ...  ...
8119             0          0          11            0  ...
8120             0          0          11            0  ...
8121             0          0           5            0  ...
8122             0          1           0            1  ...
8123             0          0          11            0  ...

      stalk-surface-below-ring  stalk-color-above-ring  \
0                            2                       7
1                            2                       7
2                            2                       7
3                            2                       7
4                            2                       7
```

```
...                          ...                     ...
8119                           2                       5
8120                           2                       5
8121                           2                       5
8122                           1                       7
8123                           2                       5

      stalk-color-below-ring  veil-type  veil-color  ring-number
ring-type  \
0                            7          0           2            1
4
1                            7          0           2            1
4
2                            7          0           2            1
4
3                            7          0           2            1
4
4                            7          0           2            1
0
...                        ...        ...         ...          ...
...
8119                         5          0           1            1
4
8120                         5          0           0            1
4
8121                         5          0           1            1
4
8122                         7          0           2            1
0
8123                         5          0           1            1
4

      spore-print-color  population  habitat
0                     2           3        5
1                     3           2        1
2                     3           2        3
3                     2           3        5
4                     3           0        1
...                 ...         ...      ...
8119                  0           1        2
8120                  0           4        2
8121                  0           1        2
8122                  7           4        2
8123                  4           1        2

[8124 rows x 22 columns]

Y
```

```
0       1
1       0
2       0
3       1
4       0
       ..
8119    0
8120    0
8121    0
8122    1
8123    0
Name: class, Length: 8124, dtype: int32
```

```
pd.DataFrame(X, columns=X.columns).describe()
```

```
         cap-shape  cap-surface    cap-color      bruises         odor
\
count  8124.000000  8124.000000  8124.000000  8124.000000  8124.000000

mean      3.348104     1.827671     4.504677     0.415559     4.144756

std       1.604329     1.229873     2.545821     0.492848     2.103729

min       0.000000     0.000000     0.000000     0.000000     0.000000

25%       2.000000     0.000000     3.000000     0.000000     2.000000

50%       3.000000     2.000000     4.000000     0.000000     5.000000

75%       5.000000     3.000000     8.000000     1.000000     5.000000

max       5.000000     3.000000     9.000000     1.000000     8.000000


       gill-attachment  gill-spacing    gill-size   gill-color  stalk-
shape  \
count      8124.000000   8124.000000  8124.000000  8124.000000
8124.000000
mean          0.974151      0.161497     0.309207     4.810684
0.567208
std           0.158695      0.368011     0.462195     3.540359
0.495493
min           0.000000      0.000000     0.000000     0.000000
0.000000
25%           1.000000      0.000000     0.000000     2.000000
0.000000
50%           1.000000      0.000000     0.000000     5.000000
1.000000
75%           1.000000      0.000000     1.000000     7.000000
1.000000
```

```
max           1.000000      1.000000     1.000000    11.000000
1.000000

       ...  stalk-surface-below-ring  stalk-color-above-ring  \
count  ...                8124.000000             8124.000000
mean   ...                   1.603644                5.816347
std    ...                   0.675974                1.901747
min    ...                   0.000000                0.000000
25%    ...                   1.000000                6.000000
50%    ...                   2.000000                7.000000
75%    ...                   2.000000                7.000000
max    ...                   3.000000                8.000000

       stalk-color-below-ring  veil-type   veil-color  ring-number  \
count             8124.000000     8124.0  8124.000000  8124.000000
mean                 5.794682        0.0     1.965534     1.069424
std                  1.907291        0.0     0.242669     0.271064
min                  0.000000        0.0     0.000000     0.000000
25%                  6.000000        0.0     2.000000     1.000000
50%                  7.000000        0.0     2.000000     1.000000
75%                  7.000000        0.0     2.000000     1.000000
max                  8.000000        0.0     3.000000     2.000000

         ring-type  spore-print-color   population      habitat
count  8124.000000        8124.000000  8124.000000  8124.000000
mean      2.291974           3.596750     3.644018     1.508616
std       1.801672           2.382663     1.252082     1.719975
min       0.000000           0.000000     0.000000     0.000000
25%       0.000000           2.000000     3.000000     0.000000
50%       2.000000           3.000000     4.000000     1.000000
75%       4.000000           7.000000     4.000000     2.000000
max       4.000000           8.000000     5.000000     6.000000

[8 rows x 22 columns]
```

Разделим выборку на обучающую и тестовую:

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
test_size=0.25, random_state=1)
print('{}, {}'.format(X_train.shape, X_test.shape))
print('{}, {}'.format(Y_train.shape, Y_test.shape))
```

```
(6093, 22), (2031, 22)
(6093,), (2031,)
```

Обучение модели

Линейная

```
SGD = SGDClassifier(max_iter=10000)
SGD.fit(X_train, Y_train)
```

```
SGDClassifier(max_iter=10000)
```

SVC

```
SVC = SVC(kernel='rbf')
SVC.fit(X_train, Y_train)
```

```
SVC()
```

```
f1_score(Y_test, SVC.predict(X_test), average='micro')
precision_score(Y_test, SVC.predict(X_test), average='micro')
```

```
0.9862136878385032
```

Дерево решений

```
DT = DecisionTreeClassifier(random_state=1)
DT.fit(X_train, Y_train)
```

```
DecisionTreeClassifier(random_state=1)
```

```
print(f1_score(Y_test, DT.predict(X_test), average='micro'))
precision_score(Y_test, DT.predict(X_test), average='micro')
```

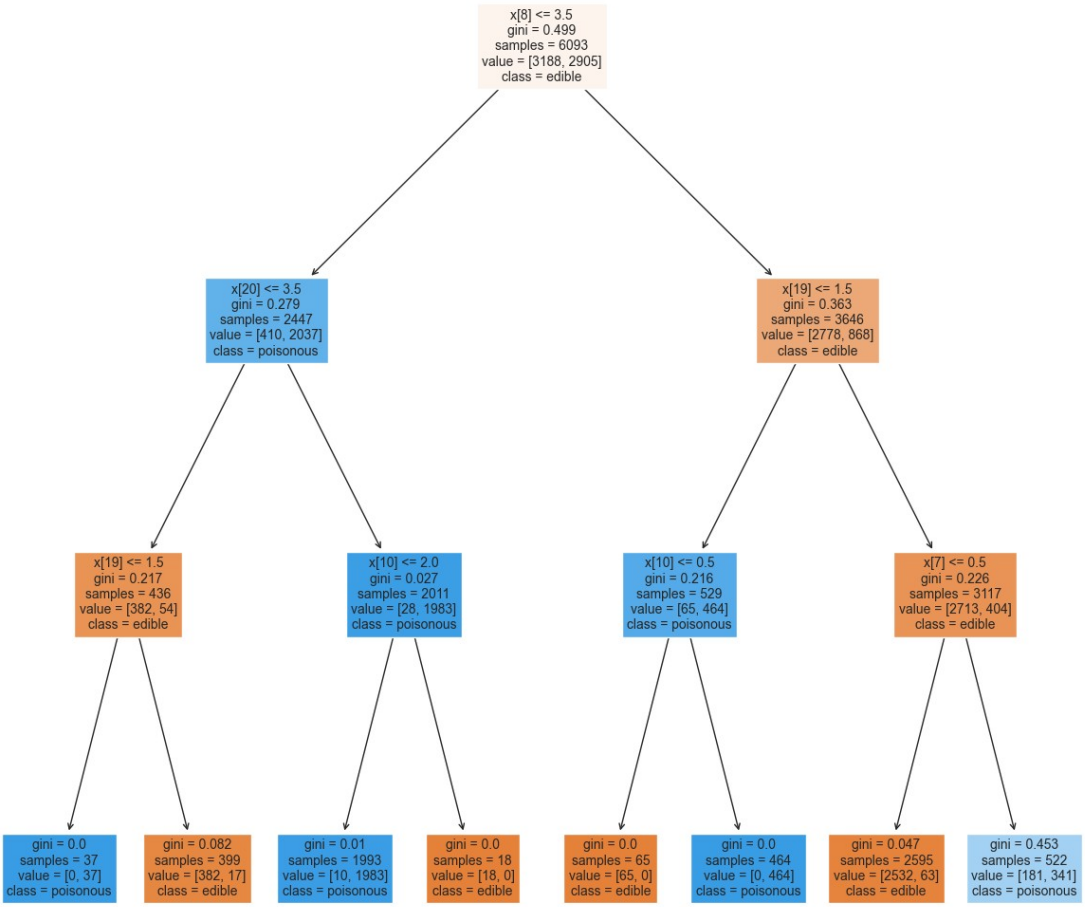
```
1.0
```

```
1.0
```

Делаем вывод, что дерево решений дает лучший результат

Визуализация

```
from sklearn import tree
fig, ax = plt.subplots(figsize=(15, 15))
clf = DecisionTreeClassifier(max_depth = 3,
                             random_state = 0)
clf.fit(X_train, Y_train)
cn=['edible', 'poisonous']
tree.plot_tree(clf, fontsize=10, class_names=cn, filled=True)
plt.show()
```


x[8] <= 3.5
gini = 0.499
samples = 6093
value = [3188, 2905]
class = edible
x[20] <= 3.5
gini = 0.279
samples = 2447
value = [410, 2037]
class = poisonous
x[19] <= 1.5
gini = 0.363
samples = 3646
value = [2778, 868]
class = edible
x[19] <= 1.5
gini = 0.217
samples = 436
value = [382, 54]
class = edible
x[10] <= 2.0
gini = 0.027
samples = 2011
value = [28, 1983]
class = poisonous
x[10] <= 0.5
gini = 0.216
samples = 529
value = [65, 464]
class = poisonous
x[7] <= 0.5
gini = 0.226
samples = 3117
value = [2713, 404]
class = edible
gini = 0.0
samples = 37
value = [0, 37]
class = poisonous
gini = 0.082
samples = 399
value = [382, 17]
class = edible
gini = 0.01
samples = 1993
value = [10, 1983]
class = poisonous
gini = 0.0
samples = 18
value = [18, 0]
class = edible
gini = 0.0
samples = 65
value = [65, 0]
class = edible
gini = 0.0
samples = 464
value = [0, 464]
class = poisonous
gini = 0.047
samples = 2595
value = [2532, 63]
class = edible
gini = 0.453
samples = 522
value = [181, 341]
class = poisonous