Ансамбли моделей машинного обучения

Цель лабораторной работы: изучение ансамблей моделей машинного обучения. Задание:

1. Выберите набор данных (датасет) для решения задачи классификации или регресии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train_test_split разделите выборку на обучающую и тестовую.
4. Обучите две ансамблевые модели. Оцените качество моделей с помощью одной из подходящих для задачи метрик. Сравните качество полученных моделей.
5. Произведите для каждой модели подбор значений одного гиперпараметра. В зависимости от используемой библиотеки можно применять функцию GridSearchCV, использовать перебор параметров в цикле, или использовать другие методы.
6. Повторите пункт 4 для найденных оптимальных значений гиперпараметров. Сравните качество полученных моделей с качеством моделей, полученных в пункте 4.

1. Подготовка данных; датасет - [https://www.kaggle.com/ronitf/heart-disease-uci/version/1 (https://www.kaggle.com/ronitf/heart-disease-uci/version/1)](https://www.kaggle.com/ronitf/heart-disease-uci/version/1)
2. age;---возраст;
3. sex;---пол;
4. chest pain type (4 values);---Тип боли;
5. resting blood pressure;---Кровяное давление в покое;
6. serum cholestoral in mg/dl;---Холестерин;
7. fasting blood sugar > 120 mg/dl;---Сахар в крови;
8. resting electrocardiographic results (values 0,1,2);---Электрокардиография в покое;
9. maximum heart rate achieved;---Максимальный сердечный ритм;
10. exercise induced angina;---Стенокардия вызванная физической нагрузкой;
11. oldpeak = ST depression induced by exercise relative to rest;---депрессия вызванная физ упражнениями;
12. the slope of the peak exercise ST segment;---Наклон пика упражнений;
13. number of major vessels (0-3) colored by flourosopy;---Кол-во крупных сосоудов по цвету thal: 3 = normal; 6 = fixed defect; 7 = reversable defect;

In [1]:

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

In [2]:

```
data = pd.read_csv('C:/Users/VTsapiy/Desktop/data/heart.csv')
```

In [3]:

```
data.head()
```

Out[3]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

In [4]:

```
data.shape
```

Out[4]:

```
(303, 14)
```

In [5]:

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

Out[5]:

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Датасет без пустых значений

Feature Scaling

In [20]:

```
from sklearn.preprocessing import MinMaxScaler
import warnings

from sklearn import svm
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_curve, auc
import pylab as pl
import matplotlib.pyplot as plt

warnings.filterwarnings('ignore')

# Create the scaler object with a range of 0-1
scaler = MinMaxScaler(feature_range=(0, 1))
# Fit on data, transform  data
scaler.fit_transform(data)
```

Out[20]:

```
array([[0.70833333, 1.        , 1.        , ..., 0.        , 0.33333333,
        1.        ],
       [0.16666667, 1.        , 0.66666667, ..., 0.        , 0.66666667,
        1.        ],
       [0.25      , 0.        , 0.33333333, ..., 0.        , 0.66666667,
        1.        ],
       ...,
       [0.8125    , 1.        , 0.        , ..., 0.5       , 1.        ,
        0.        ],
       [0.58333333, 1.        , 0.        , ..., 0.25      , 1.        ,
        0.        ],
       [0.58333333, 0.        , 0.33333333, ..., 0.25      , 0.66666667,
        0.        ]])
```

Разделим датасет на тестовую и обучающую выборки

In [8]:

```
X = data.drop('target',axis = 1).values
y = data['target'].values
```

Ансамблевые модели

In [37]:

```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import ExtraTreesClassifier, ExtraTreesRegressor

from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import balanced_accuracy_score
from sklearn.metrics import precision_score, recall_score, f1_score

from sklearn.model_selection import GridSearchCV
```

In [12]:

```
kfold = 5 #количество подвыборок для валидации
```

In [13]:

```
itog_val = {} #список для записи результатов кросс валидации разных алгоритмов
```

In [14]:

```
ROCtrainTRN, ROCtestTRN, ROCtrainTRG, ROCtestTRG = train_test_split(X, y, test_size=0.20)
```

In [15]:

```
model_rfc = RandomForestClassifier(n_estimators = 75) #в параметре передаем кол-во деревьев
model_knc = KNeighborsClassifier(n_neighbors = 20) #в параметре передаем кол-во соседей
model_lr = LogisticRegression(penalty='l1', tol=0.01)
model_svc = svm.SVC() #по умолчанию kernek='rbf'
```

1. SVM - метод опорных векторов(SVC)
2. Метод k-ближайших соседей(KNeighborsClassifier)
3. Random forest(RandomForestClassifier)
4. Логистическая регрессия (LogisticRegression)

In [19]:

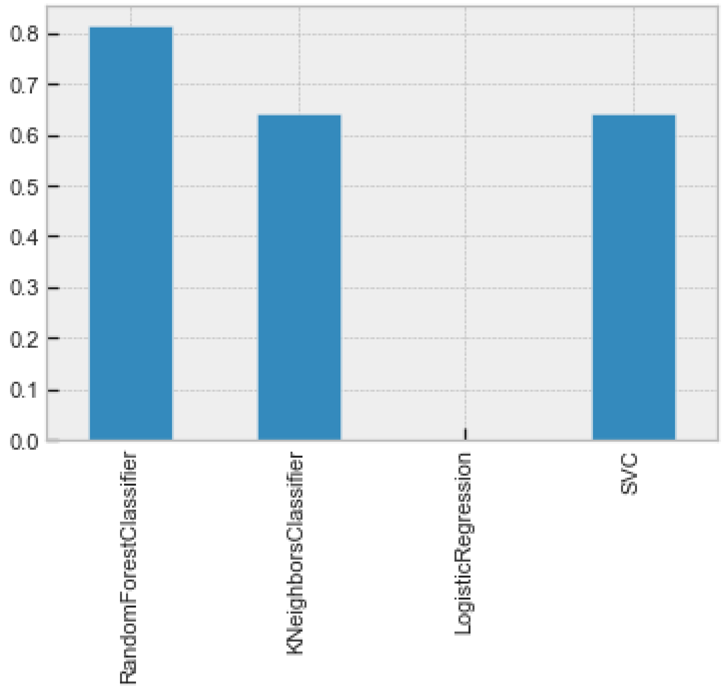
```
scores = cross_val_score(model_rfc, X, y, cv = kfold)
itog_val['RandomForestClassifier'] = scores.mean()
scores = cross_val_score(model_knc, X, y, cv = kfold)
itog_val['KNeighborsClassifier'] = scores.mean()
scores = cross_val_score(model_lr, X, y, cv = kfold)
itog_val['LogisticRegression'] = scores.mean()
scores = cross_val_score(model_svc, X, y, cv = kfold)
itog_val['SVC'] = scores.mean()
```

In [21]:

```
plt.style.use('bmh')
data.from_dict(data = itog_val, orient='index').plot(kind='bar', legend=False)
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xd2e22b0>
```



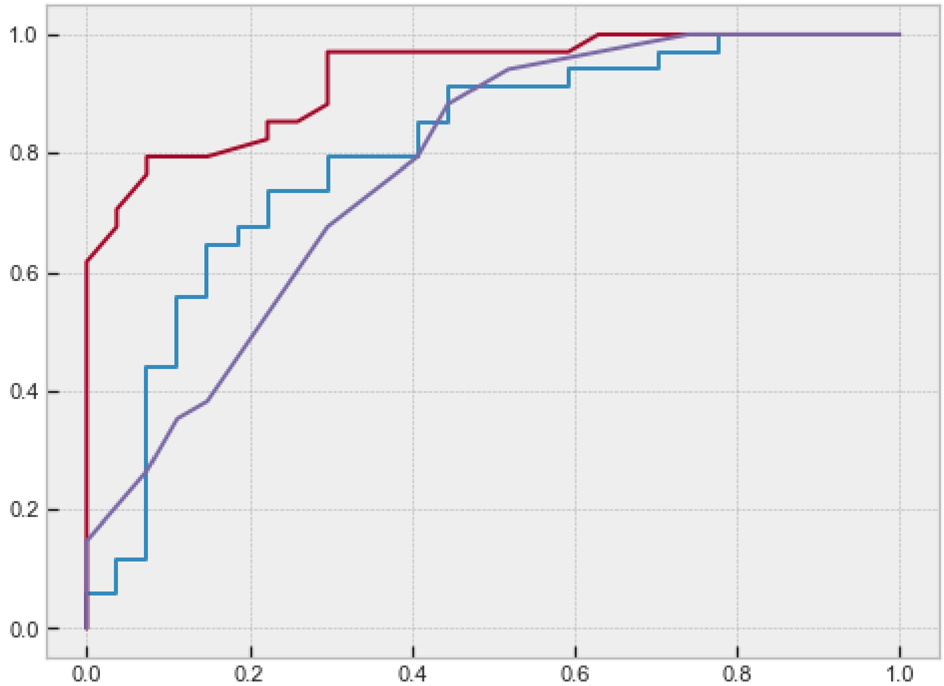
In [22]:

```
pl.clf()
plt.figure(figsize=(8,6))
#SVC
model_svc.probability = True
probas = model_svc.fit(ROCtrainTRN, ROCtrainTRG).predict_proba(ROCtestTRN)
fpr, tpr, thresholds = roc_curve(ROCtestTRG, probas[:, 1])
roc_auc  = auc(fpr, tpr)
pl.plot(fpr, tpr, label='%s ROC (area = %0.2f)' % ('SVC', roc_auc))
#RandomForestClassifier
probas = model_rfc.fit(ROCtrainTRN, ROCtrainTRG).predict_proba(ROCtestTRN)
fpr, tpr, thresholds = roc_curve(ROCtestTRG, probas[:, 1])
roc_auc  = auc(fpr, tpr)
pl.plot(fpr, tpr, label='%s ROC (area = %0.2f)' % ('RandonForest',roc_auc))
#KNeighborsClassifier
probas = model_knc.fit(ROCtrainTRN, ROCtrainTRG).predict_proba(ROCtestTRN)
fpr, tpr, thresholds = roc_curve(ROCtestTRG, probas[:, 1])
roc_auc  = auc(fpr, tpr)
pl.plot(fpr, tpr, label='%s ROC (area = %0.2f)' % ('KNeighborsClassifier',roc_auc))
#LogisticRegression
probas = model_lr.fit(ROCtrainTRN, ROCtrainTRG).predict_proba(ROCtestTRN)
fpr, tpr, thresholds = roc_curve(ROCtestTRG, probas[:, 1])
roc_auc  = auc(fpr, tpr)
pl.plot(fpr, tpr, label='%s ROC (area = %0.2f)' % ('LogisticRegression',roc_auc))
pl.plot([0, 1], [0, 1], 'k--')
pl.xlim([-0.1, 1.1])
pl.ylim([-0.1, 1.1])
pl.xlabel('False Positive Rate')
pl.ylabel('True Positive Rate')
pl.legend(loc=0, fontsize='small')
pl.show()
```

```
---------------------------------------------------------------------------
-
ValueError                                Traceback (most recent call las
t)
<ipython-input-22-42a13da60fc0> in <module>
     18 pl.plot(fpr, tpr, label='%s ROC (area = %0.2f)' % ('KNeighborsClas
sifier',roc_auc))
     19 #LogisticRegression
---> 20 probas = model_lr.fit(ROCtrainTRN, ROCtrainTRG).predict_proba(ROCt
estTRN)
     21 fpr, tpr, thresholds = roc_curve(ROCtestTRG, probas[:, 1])
     22 roc_auc  = auc(fpr, tpr)

c:\users\vtsapiy\appdata\local\programs\python\python37-32\lib\site-packag
es\sklearn\linear_model\_logistic.py in fit(self, X, y, sample_weight)
   1486         The SAGA solver supports both float64 and float32 bit arra
ys.
   1487         """
-> 1488         solver = _check_solver(self.solver, self.penalty, self.dua
l)
   1489 
   1490         if not isinstance(self.C, numbers.Number) or self.C < 0:

c:\users\vtsapiy\appdata\local\programs\python\python37-32\lib\site-packag
es\sklearn\linear_model\_logistic.py in _check_solver(solver, penalty, dua
l)
```

```
    443     if solver not in ['liblinear', 'saga'] and penalty not in ('l
2', 'none'):
    444         raise ValueError("Solver %s supports only 'l2' or 'none' p
enalties, "
--> 445                          "got %s penalty." % (solver, penalty))
    446     if solver != 'liblinear' and dual:
    447         raise ValueError("Solver %s supports only "

ValueError: Solver lbfgs supports only 'l2' or 'none' penalties, got l1 pe
nalty.
```

```
<Figure size 432x288 with 0 Axes>
```



In [24]:

```
# Функция train_test_split разделила исходную выборку таким образом,
#чтобы в обучающей и тестовой частях сохранились пропорции классов.
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.35, random_state=1)
```

In [25]:

```
from sklearn.preprocessing import MinMaxScaler
warnings.filterwarnings('ignore')

# Create the scaler object with a range of 0-1
scaler = MinMaxScaler(feature_range=(0, 1))
# Fit on data, transform  data
scaler.fit_transform(X)
scaler.fit_transform(X_train)
scaler.fit_transform(X_test)
```

Out[25]:

```
array([[0.77777778, 0.        , 0.        , ..., 0.        , 0.75      ,
        1.        ],
       [0.61111111, 1.        , 0.33333333, ..., 1.        , 0.        ,
        1.        ],
       [0.38888889, 1.        , 0.        , ..., 1.        , 0.5       ,
        1.        ],
       ...,
       [0.52777778, 1.        , 0.        , ..., 0.        , 0.        ,
        1.        ],
       [0.66666667, 1.        , 0.33333333, ..., 0.5       , 1.        ,
        1.        ],
       [0.38888889, 1.        , 0.33333333, ..., 0.        , 0.        ,
        1.        ]])
```

In [26]:

```
rfc = RandomForestClassifier().fit(X_train, y_train)
predicted_rfc = rfc.predict(X_test)
```

In [27]:

```
accuracy_score(y_test, predicted_rfc)
```

Out[27]:

```
0.7570093457943925
```

In [28]:

```
balanced_accuracy_score(y_test, predicted_rfc)
```

Out[28]:

```
0.7559649122807017
```

In [29]:

```
(precision_score(y_test, predicted_rfc, average='weighted'),
 recall_score(y_test, predicted_rfc, average='weighted'))
```

Out[29]:

```
(0.7570093457943925, 0.7570093457943925)
```

In [30]:

```
f1_score(y_test, predicted_rfc, average='weighted')
```

Out[30]:

```
0.7570093457943925
```

In [31]:

```
abc = AdaBoostClassifier().fit(X_train, y_train)
predicted_abc = abc.predict(X_test)
```

In [32]:

```
accuracy_score(y_test, predicted_abc)
```

Out[32]:

```
0.7289719626168224
```

In [33]:

```
balanced_accuracy_score(y_test, predicted_abc)
```

Out[33]:

```
0.7284210526315789
```

In [34]:

```
(precision_score(y_test, predicted_abc, average='weighted'),
 recall_score(y_test, predicted_abc, average='weighted'))
```

Out[34]:

```
(0.7293842770753162, 0.7289719626168224)
```

In [35]:

```
f1_score(y_test, predicted_abc, average='weighted')
```

Out[35]:

```
0.7291144464706996
```

In [36]:

```
rfc_n_range = np.array(range(5,100,5))
rfc_tuned_parameters = [{'n_estimators': rfc_n_range}]
rfc_tuned_parameters
```

Out[36]:

```
[{'n_estimators': array([ 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65,
70, 75, 80, 85,
         90, 95])}]
```

In [38]:

```
warnings.filterwarnings('ignore')

gs_rfc = GridSearchCV(RandomForestClassifier(), rfc_tuned_parameters, cv=5,
                      scoring='accuracy')
gs_rfc.fit(X_train, y_train)
```

Out[38]:

```
GridSearchCV(cv=5, error_score=nan,
             estimator=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,
                                              class_weight=None,
                                              criterion='gini', max_depth=No
ne,
                                              max_features='auto',
                                              max_leaf_nodes=None,
                                              max_samples=None,
                                              min_impurity_decrease=0.0,
                                              min_impurity_split=None,
                                              min_samples_leaf=1,
                                              min_samples_split=2,
                                              min_weight_fraction_leaf=0.0,
                                              n_estimators=100, n_jobs=None,
                                              oob_score=False,
                                              random_state=None, verbose=0,
                                              warm_start=False),
             iid='deprecated', n_jobs=None,
             param_grid=[{'n_estimators': array([ 5, 10, 15, 20, 25, 30, 35,
40, 45, 50, 55, 60, 65, 70, 75, 80, 85,
       90, 95])}],
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring='accuracy', verbose=0)
```

In [39]:

```
gs_rfc.best_params_
```

Out[39]:

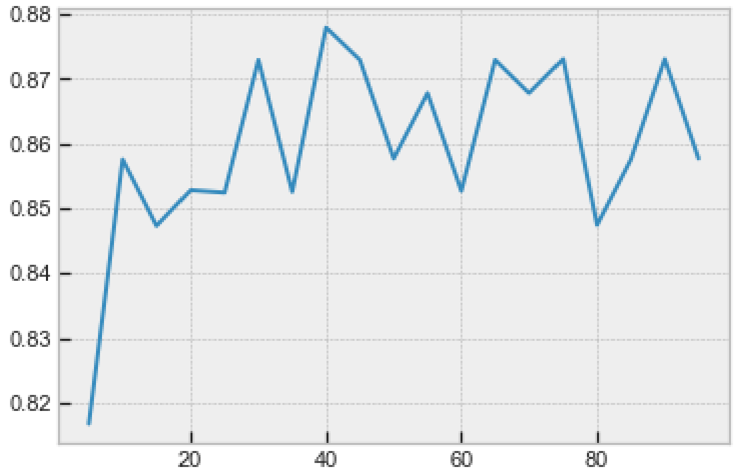
```
{'n_estimators': 40}
```

In [40]:

```
plt.plot(rfc_n_range, gs_rfc.cv_results_['mean_test_score'])
```

Out[40]:

```
[<matplotlib.lines.Line2D at 0x4ed58f0>]
```



In [41]:

```
abc_n_range = np.array(range(5,100,5))
abc_tuned_parameters = [{'n_estimators': abc_n_range}]
abc_tuned_parameters
```

Out[41]:

```
[{'n_estimators': array([ 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65,
70, 75, 80, 85,
         90, 95])}]
```

In [42]:

```
gs_abc = GridSearchCV(AdaBoostClassifier(), abc_tuned_parameters, cv=5,
                      scoring='accuracy')
gs_abc.fit(X_train, y_train)
```

Out[42]:

```
GridSearchCV(cv=5, error_score=nan,
             estimator=AdaBoostClassifier(algorithm='SAMME.R',
                                          base_estimator=None,
                                          learning_rate=1.0, n_estimators=5
0,
                                          random_state=None),
             iid='deprecated', n_jobs=None,
             param_grid=[{'n_estimators': array([ 5, 10, 15, 20, 25, 30, 35,
40, 45, 50, 55, 60, 65, 70, 75, 80, 85,
       90, 95])}],
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring='accuracy', verbose=0)
```

In [43]:

```
gs_abc.best_params_
```

Out[43]:

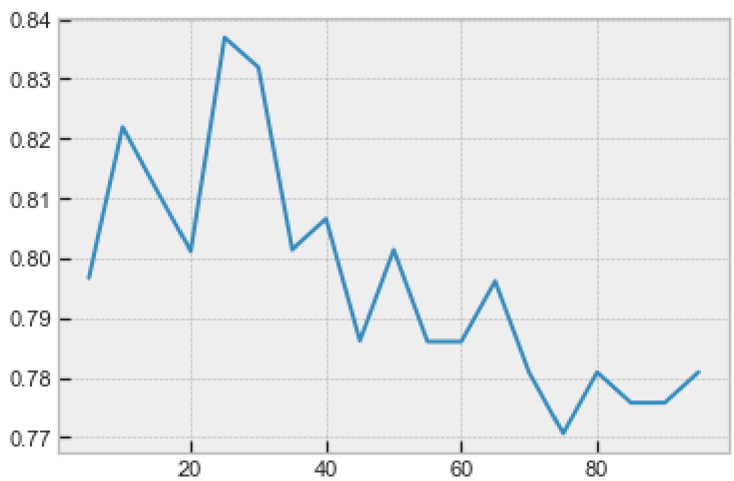
```
{'n_estimators': 25}
```

In [44]:

```
plt.plot(abc_n_range, gs_abc.cv_results_['mean_test_score'])
```

Out[44]:

```
[<matplotlib.lines.Line2D at 0x4f274f0>]
```



In [45]:

```
rfc_optimized = RandomForestClassifier(n_estimators=gs_rfc.best_params_['n_estimators']).fi
predicted_rfc_opt = rfc_optimized.predict(X_test)
```

In [46]:

```
accuracy_score(y_test, predicted_rfc_opt)
```

Out[46]:

```
0.7289719626168224
```

In [47]:

```
balanced_accuracy_score(y_test, predicted_rfc_opt)
```

Out[47]:

```
0.7284210526315789
```

In [48]:

```
(precision_score(y_test, predicted_rfc_opt, average='weighted'),
 recall_score(y_test, predicted_rfc_opt, average='weighted'))
```

Out[48]:

```
(0.7293842770753162, 0.7289719626168224)
```

In [49]:

```
f1_score(y_test, predicted_rfc_opt, average='weighted')
```

Out[49]:

```
0.7291144464706996
```

In [50]:

```
abc_optimized = RandomForestClassifier(n_estimators=gs_abc.best_params_['n_estimators']).fi
predicted_abc_opt = abc_optimized.predict(X_test)
```

In [51]:

```
accuracy_score(y_test, predicted_abc_opt)
```

Out[51]:

```
0.7289719626168224
```

In [52]:

```
balanced_accuracy_score(y_test, predicted_abc_opt)
```

Out[52]:

```
0.7271929824561403
```

In [53]:

```
(precision_score(y_test, predicted_abc_opt, average='weighted'),
 recall_score(y_test, predicted_abc_opt, average='weighted'))
```

Out[53]:

```
(0.7287187514387, 0.7289719626168224)
```

In [54]:

```
f1_score(y_test, predicted_abc_opt, average='weighted')
```

Out[54]:

```
0.7287815169164215
```

Сравнивая модели, можно сделать вывод что все методы примерно одинаковые

In [ ]:

```

```