1. Описание базы данных

`Набор данных по распознаванию вин`

1. `Alcohol` - Алкоголь
2. `Malic acid` - Яблочная кислота
3. `Ash` - Пепел
4. `Alcalinity of ash` - Щелочность золы
5. `Magnesium` - Магний
6. `Total phenols` - Общие фенолы
7. `Flavanoids` - Флаваноиды
8. `Nonflavanoid phenols` - Нефлаваноидные фенолы
9. `Proanthocyanins` - Проантоцианы
10. `Color intensity` - Интенсивность цвета
11. `Hue` - Оттенок
12. `OD280/OD315 of diluted wines` - OD280 / OD315 разбавленных вин
13. `Proline` - Пролин

РК 1-4-3

`Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?`

In [78]:
```
from operator import itemgetter
import matplotlib. pyplot as plt
import matplotlib. ticker as ticker
import numpy as np
import pandas as pd
import pandas_profiling
import math
from simple_kNN import *
```

Подготовка данных и построение базовых моделей для оценки качества

In [79]:
```
from sklearn.datasets import *
from typing import Dict, Tuple
from scipy import stats
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsRegressor, KNeighborsClassifier
from sklearn.metrics import accuracy_score, balanced_accuracy_score
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import precision_score, recall_score, f1_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.metrics import roc_curve, roc_auc_score, classification_report
import seaborn as sns
import matplotlib. pyplot as plt
%matplotlib inline
sns. set(style="ticks")
```

In [80]:
```
wine = load_wine()
```

In [81]:
```
wine.feature_names
```

Out[81]:
```
['alcohol',
 'malic_acid',
 'ash',
 'alcalinity_of_ash',
 'magnesium',
 'total_phenols',
 'flavanoids',
 'nonflavanoid_phenols',
 'proanthocyanins',
 'color_intensity',
 'hue',
 'od280/od315_of_diluted_wines',
 'proline']
```

In [82]:
```
np.unique(wine.target)
```

Out[82]:
```
array([0, 1, 2])
```

In [83]:
```
wine.target_names
```

Out[83]:
```
array(['class_0', 'class_1', 'class_2'], dtype='<U7')
```

In [84]:
```
list(zip(np.unique(wine.target), wine.target_names))
```

Out[84]:
```
[(0, 'class_0'), (1, 'class_1'), (2, 'class_2')]
```

In [85]:
```
wine.data.shape, wine.target.shape
```

Out[85]:
```
((178, 13), (178,))
```

In [86]:
```
# Сформируем DataFrame
wine_df = pd.DataFrame(data=np.c_[wine['data'], wine['target']], columns=wine['feature_names'] + ['target'])
```

In [87]:
```
wine_df
```

Out[87]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065.0	0.0
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050.0	0.0
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185.0	0.0
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480.0	0.0
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
173	13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	7.70	0.64	1.74	740.0	2.0
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	0.43	1.41	7.30	0.70	1.56	750.0	2.0
175	13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	10.20	0.59	1.56	835.0	2.0
176	13.17	2.59	2.37	20.0	120.0	1.65	0.68	0.53	1.46	9.30	0.60	1.62	840.0	2.0
177	14.13	4.10	2.74	24.5	96.0	2.05	0.76	0.56	1.35	9.20	0.61	1.60	560.0	2.0

178 rows × 14 columns

In [88]:
```
wine_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   alcohol                       178 non-null    float64
 1   malic_acid                    178 non-null    float64
 2   ash                           178 non-null    float64
 3   alcalinity_of_ash             178 non-null    float64
 4   magnesium                     178 non-null    float64
 5   total_phenols                 178 non-null    float64
 6   flavanoids                    178 non-null    float64
 7   nonflavanoid_phenols          178 non-null    float64
 8   proanthocyanins               178 non-null    float64
 9   color_intensity               178 non-null    float64
 10  hue                           178 non-null    float64
 11  od280/od315_of_diluted_wines  178 non-null    float64
 12  proline                       178 non-null    float64
 13  target                        178 non-null    float64
dtypes: float64(14)
memory usage: 19.6 KB
```

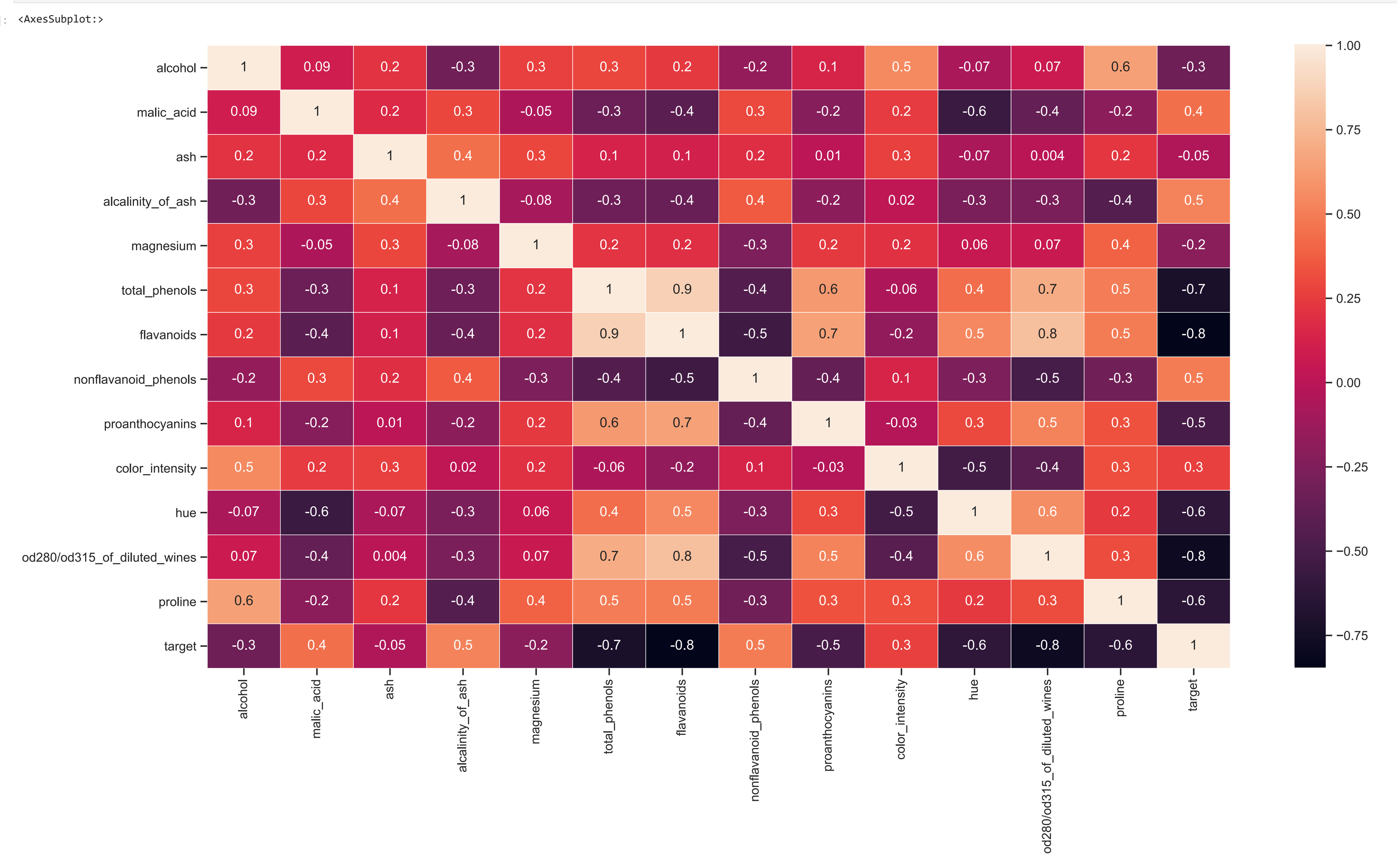
In [89]:
```
wine_df.describe()
```

Out[89]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.058090	0.957449	2.611685	746.893258	0.938202
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318286	0.228572	0.709990	314.907474	0.775035
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280000	0.480000	1.270000	278.000000	0.000000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220000	0.782500	1.937500	500.500000	0.000000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690000	0.965000	2.780000	673.500000	1.000000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200000	1.120000	3.170000	985.000000	2.000000
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.000000	1.710000	4.000000	1680.000000	2.000000

In [90]:
```
fig, ax = plt.subplots(figsize=(20,10))
sns.heatmap(wine_df.corr(), annot = True, fmt='.1g', linewidths=.5, ax=ax)
```

Out[90]:
```
<AxesSubplot:>
```



In [91]:
```
# pandas_profiling.ProfileReport(wine_df)
```

```
Summarize dataset: 100%|██████████| 27/27 [00:18<00:00,  1.45it/s, Completed]
Generate report structure: 100%|██████████| 1/1 [00:04<00:00,  4.39s/it]
Render HTML: 100%|██████████| 1/1 [00:02<00:00,  2.14s/it]
```

Out[91]:

Разделение выборки на обучающую и тестовую

In [92]:
```
wine_X_train, wine_X_test, wine_y_train, wine_y_test = train_test_split(wine.data, wine.target, test_size=0.5, random_state=1)
```

In [93]:
```
wine_X_train.shape, wine_y_train.shape
```

Out[93]:
```
((89, 13), (89,))
```

In [94]:
```
wine_X_test.shape, wine_y_test.shape
```

Out[94]:
```
((89, 13), (89,))
```

In [95]:
```
np.unique(wine_y_train)
```

Out[95]:
```
array([0, 1, 2])
```

In [96]:
```
np.unique(wine_y_test)
```

Out[96]:
```
array([0, 1, 2])
```

Построение базовой модели на основе ближайших соседей

In [97]:
```
# 2 ближайших соседа
cl1_1 = KNeighborsClassifier(n_neighbors=2)
cl1_1.fit(wine_X_train, wine_y_train)
target1_1 = cl1_1.predict(wine_X_test)
len(target1_1), target1_1
```

Out[97]:
```
(89,
 array([0, 1, 2, 1, 0, 1, 2, 0, 1, 1, 0, 1, 1, 0, 2, 1, 1, 1, 1, 0, 0, 1,
        1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0,
        0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 2, 0,
        1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 2, 2, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0,
        1]))
```

In [98]:
```
# 10 ближайших соседа
cl1_2 = KNeighborsClassifier(n_neighbors=10)
cl1_2.fit(wine_X_train, wine_y_train)
target1_2 = cl1_2.predict(wine_X_test)
len(target1_2), target1_2
```

Out[98]:
```
(89,
 array([1, 1, 2, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1,
        2, 0, 2, 1, 0, 0, 0, 2, 1, 2, 1, 0, 2, 1, 1, 1, 1, 0, 0, 1, 1, 2,
        0, 2, 2, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 2, 2, 1, 0,
        1, 0, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1, 0, 2, 1, 0, 1, 0, 1, 1, 1, 0,
        1]))
```

Метрики качества классификации

In [99]:
```
accuracy_score(wine_y_test, target1_1)
```

Out[99]:
```
0.6404494382022472
```

In [100…
```
accuracy_score(wine_y_test, target1_2)
```

Out[100…
```
0.6629213483146067
```

Произведите подбор гиперпараметра K с использованием GridSearchCV

In [101…
```
print(classification_report(wine_y_test, target1_2))
```

```
              precision    recall  f1-score   support

           0       0.93      0.79      0.85        33
           1       0.61      0.82      0.70        34
           2       0.33      0.23      0.27        22

    accuracy                           0.66        89
   macro avg       0.62      0.61      0.61        89
weighted avg       0.66      0.66      0.65        89
```

In [102…
```
def print_gridResults(grid):
    print(f'Подобранный параметр: {grid.best_params_}')
    print(f'Оценка при подобранном параметре: {grid.best_score_}')
    return [grid.best_params_, grid.best_score_]
```

In [103…
```
param_grid = {'n_neighbors' : np.arange(1, 25)}
grid = GridSearchCV(KNeighborsClassifier(), param_grid, cv=5)
grid.fit(wine_X_train, wine_y_train)
print_gridResults(grid)
```

```
Подобранный параметр: {'n_neighbors': 17}
Оценка при подобранном параметре: 0.7405228758169934
```

Out[103…
```
[{'n_neighbors': 17}, 0.7405228758169934]
```

Произведите подбор гиперпараметра K с использованием RandomizedSearchCV

In [104…
```
param_grid = {'n_neighbors' : np.arange(1, 25)}
grid2 = RandomizedSearchCV(KNeighborsClassifier(), param_grid, cv=5, n_iter=2)
grid2.fit(wine_X_train, wine_y_train)
print_gridResults(grid2)
```

```
Подобранный параметр: {'n_neighbors': 15}
Оценка при подобранном параметре: 0.7189542483660132
```

Out[104…
```
[{'n_neighbors': 15}, 0.7189542483660132]
```

Кросс валидация

In [105…
```
DEFAULT_FOLDS = 3
DEFAULT_RANGES = np.arange(1, 25)
```

In [106…
```
cv1 = cross_val_score(KNeighborsClassifier(n_neighbors=5), wine.data, wine.target, cv=DEFAULT_FOLDS, scoring='accuracy')
```

In [107…
```
cv1
```

Out[107…
```
array([0.61666667, 0.61016949, 0.76271186])
```

In [108…
```
from sklearn.model_selection import *
```

In [109…
```
def knn_cross_strtegies(mass_strategies, mass_matrics, k_range, cv, x, y):
    k_strategies_scores = {}
    for strat in mass_strategies:
        strategia = strat(cv) # n_splits
        k_scores = knn_cross_val_score(mass_matrics, k_range, strategia, x, y)
        k_strategies_scores[strat.__name__] = k_scores

    return k_strategies_scores

def knn_cross_val_score(mass_matrics, k_range, cv, x, y):
    k_scores = {}
    for metric in mass_matrics:
        for k in k_range:
            knn = KNeighborsClassifier(n_neighbors=k)
            scores = cross_val_score(knn, x, y, cv=cv, scoring=metric)
            if metric not in k_scores:
                k_scores.update({metric : []})

            k_scores[metric].append(scores.mean())
    return k_scores

def plot_knn_cross_val_score(k_scores, cv):
    columns=len(k_scores.items())
    fig = plt.figure(figsize=(10*columns, 10))
    rows=1
    index = 1
    for stratName, metrics in k_scores.items():
        ax = fig.add_subplot(rows, columns, index)
        for name, scores in metrics.items():
            ax.plot(np.arange(1, len(scores)+1), scores)
        ax.set_title(stratName)
        ax.legend(metrics.keys(), loc='lower right')
        index+=1
```

In [110…
```
mass_metrics = ['accuracy', 'balanced_accuracy', 'f1_micro', 'f1_macro']
strategies = [KFold, RepeatedKFold, StratifiedKFold, RepeatedStratifiedKFold, ShuffleSplit, StratifiedShuffleSplit, TimeSeriesSplit] #
```

In [111…
```
scores = knn_cross_strtegies(strategies, mass_metrics, DEFAULT_RANGES, DEFAULT_FOLDS, wine.data, wine.target)
```

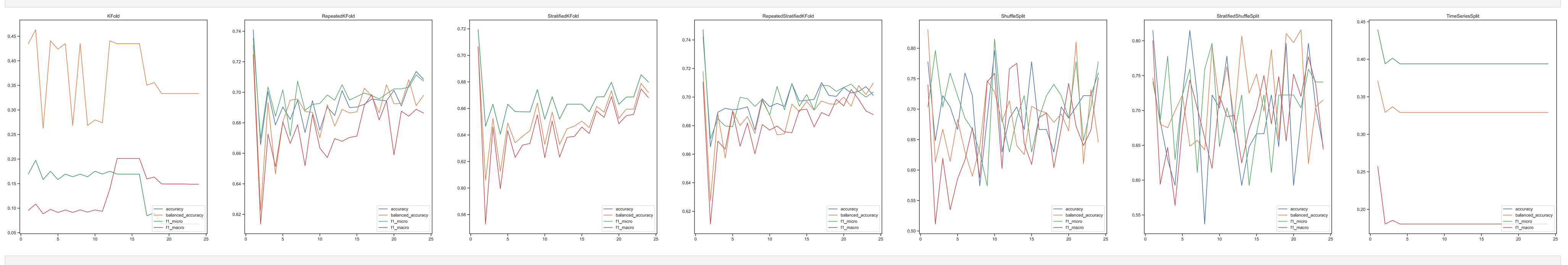
In [112…
```
# scores
```

In [113…
```
# scores, range = knn_cross_val_score(mass_metrics, np.arange(1,25), DEFAULT_FOLDS, wine.data, wine.target)
# scores, range
```

Сравнение моделей

In [114…
```
# plot_knn_cross_val_score(scores, k_range)
```

In [115…
```
plot_knn_cross_val_score(scores, DEFAULT_FOLDS)
```



In [ ]: