МИНЕСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
Кафедра «Измерительно-вычислительные комплексы»

«Технология обработки информации»

Отчет по лабораторной работе №5

Исследование инструментов классификации библиотеки Scikit-learn

Выполнил:
Студент группы ИСТбд-41
Калашников М. А.
Проверил:
Шишкин В.В.

Ульяновск
2022

1. Ознакомиться с классификаторами библиотеки Scikit-learn
2. Выбрать для исследования не менее 3 классификаторов

Здравствуйте, хочу утвердить 3 классификатора и набор данных для машинного обучения. Для 5 лабораторной работы.
Классификаторы: 1) KNN KNeighborsClassifier; 2) Древо решений DecisionTreeClassifier; 3) Логистическая регрессия LogisticRegression

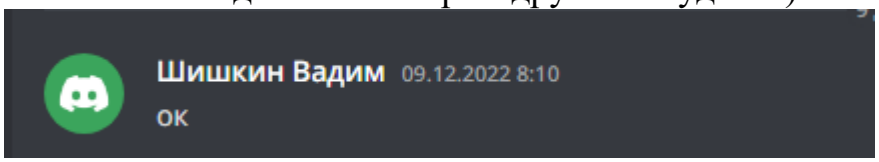
3. Выбрать набор данных для задач классификации из открытых источников
https://tproger.ru/translations/the-best-datasets-for-machine-learning-and-data-science/
https://vc.ru/ml/150241-15-proektov-dlya-razvitiya-navykov-raboty-s-mashinnym-obucheniem
https://archive.ics.uci.edu/ml/index.php
https://habr.com/ru/company/edison/blog/480408/
https://www.kaggle.com/datasets/
учебные наборы библиотеки Scikit-learn

Набор данных: Red Wine Quality https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Red Wine Quality

Simple and clean practice dataset for regression or classification modelling

4. Выбор классификаторов и набора данных утвердить у преподавателя (не должно быть полного совпадения с выбором другого студента)



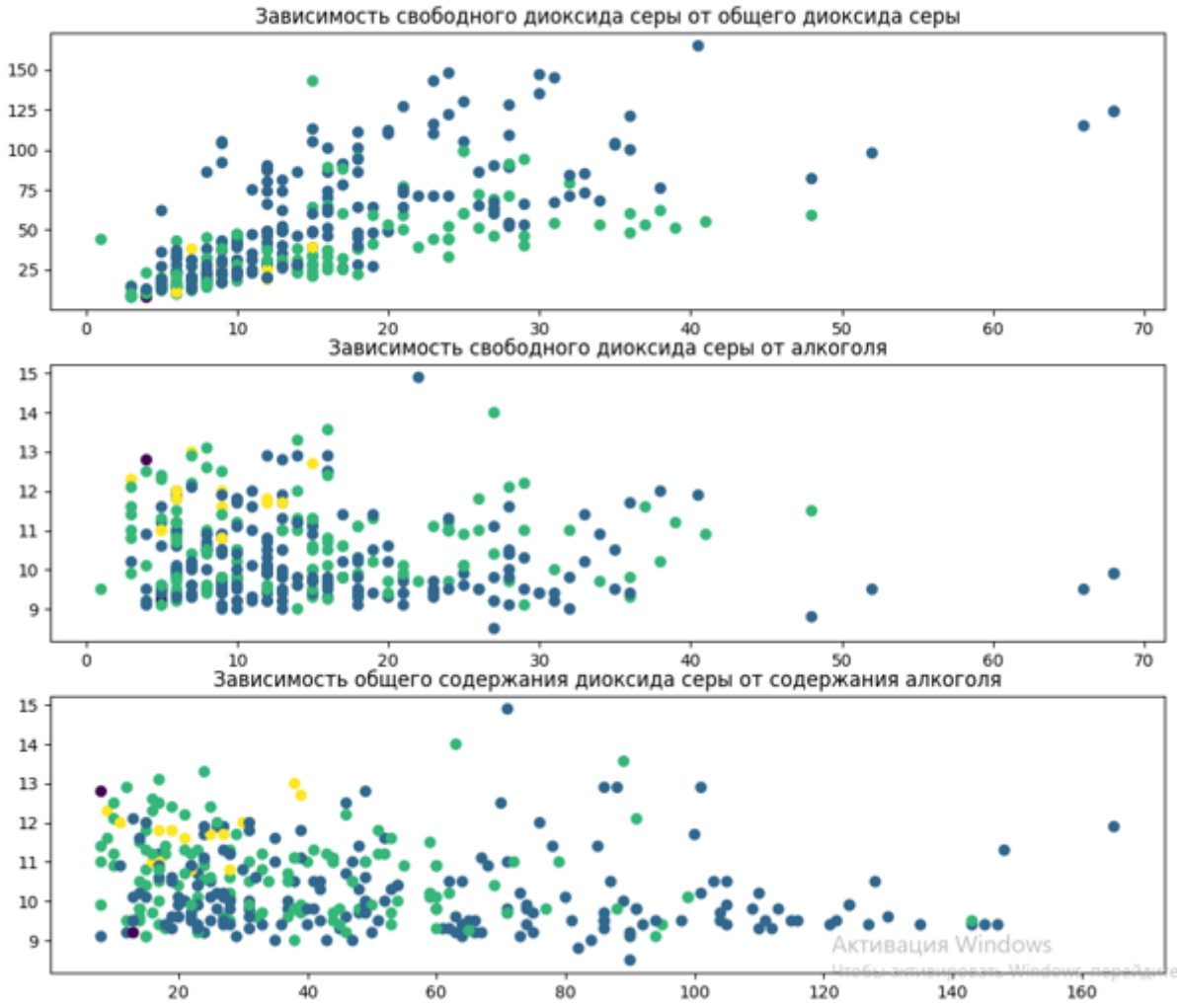
5. Для каждого классификатора определить целевой столбец и набор признаков. Обосновать свой выбор. При необходимости преобразовать типы признаковых данных.

```
# Рисует основные признаки классификации
    v1, orig_v1 = plt.subplots(3, 1, figsize=(12, 12))
    orig_v1[0].scatter(KNN_X_Test[:, 5], KNN_X_Test[:, 6], c=Y_nachal_KNN)
    orig_v1[0].set_title('Зависимость свободного диоксида серы от общего диоксида серы')
    orig_v1[1].scatter(KNN_X_Test[:, 5], KNN_X_Test[:, 10], c=Y_nachal_KNN)
    orig_v1[1].set_title('Зависимость свободного диоксида серы от алкоголя')
    orig_v1[2].scatter(KNN_X_Test[:, 6], KNN_X_Test[:, 10], c=Y_nachal_KNN)
    orig_v1[2].set_title('Зависимость общего содержания диоксида серы от содержания алкоголя')
    plt.show()
```

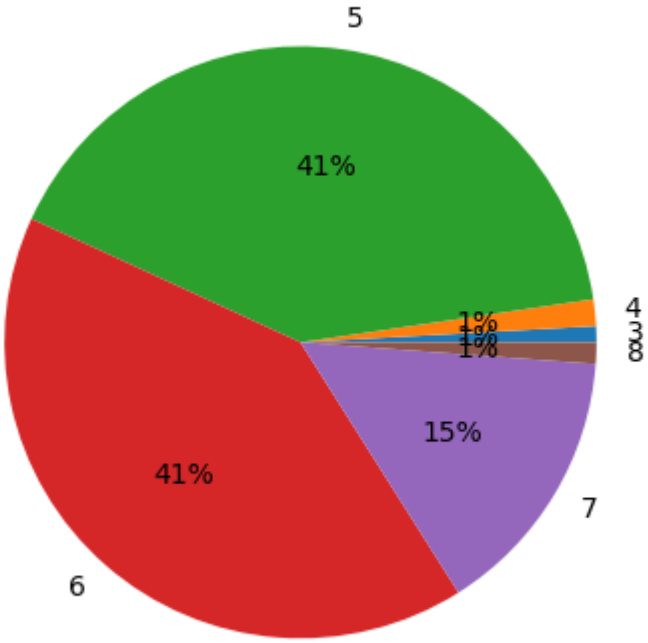
6. Подготовить данные к обучению.

```
# Набор массива данных Red Wine Quality https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009
dataset = pd.read_csv('winequality-red.csv')
# Содержимое dataset (0 - 5)
print('Просмотр пяти рядов\n', dataset.head())
# Вывод размерности
print('Формирование набора данных : ', dataset.shape)
# Количество пропусков
print('Количество отсутствующих значений во всем наборе данных\n', dataset.isnull().sum())
# Описательная статистика
print('Статистика\n', dataset.describe().round(2))
# Удаление столбца
dataset.drop(columns='Id', inplace=True)
# Уникальные значения качества (quality)
print('\nКачество ценности : ', dataset['quality'].unique())
# Группирование по значениям quality
ave_qu = dataset.groupby('quality').mean()
print('Группировка по качеству\n', ave_qu.round(2))
# Формирования набора данных
X = dataset.drop(columns='quality').values
y = dataset['quality'].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

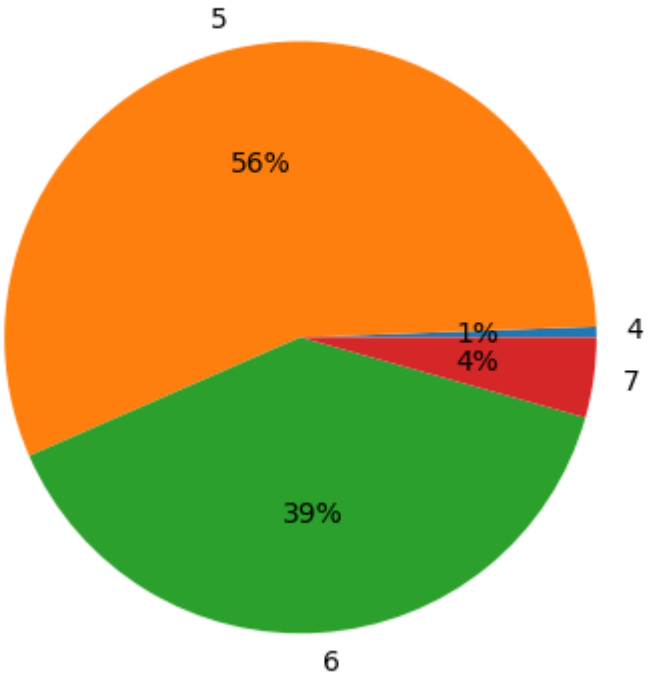
7. Провести обучение и оценку моделей на сырых данных.

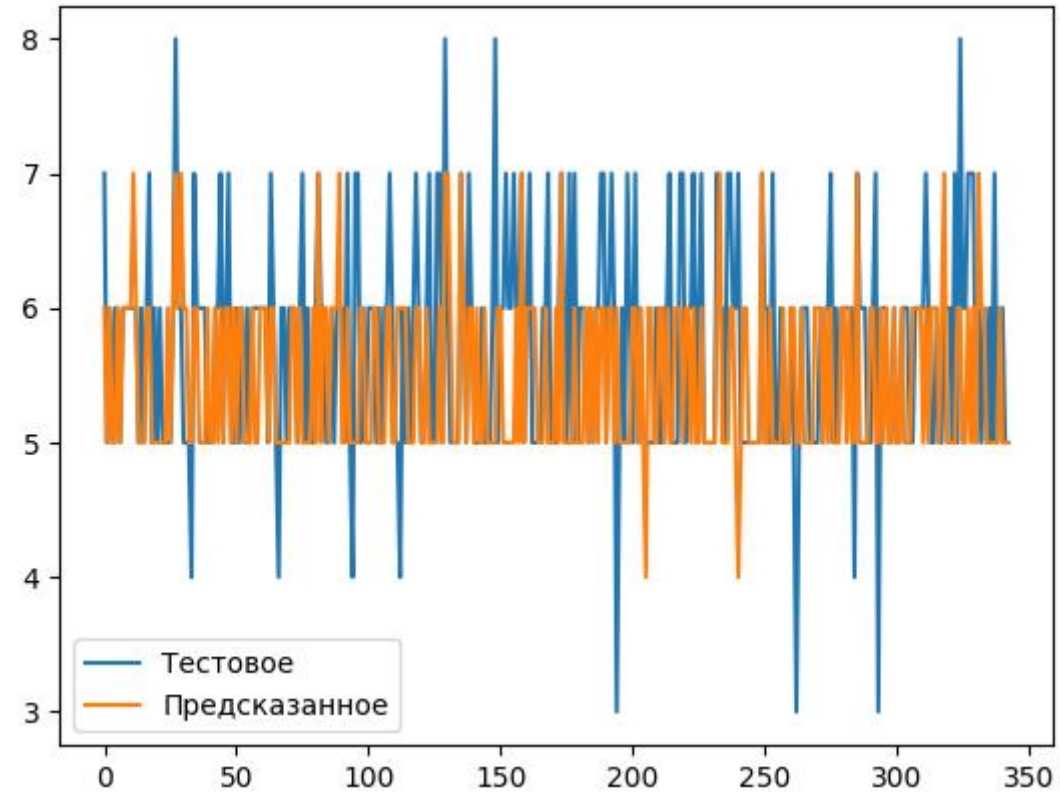


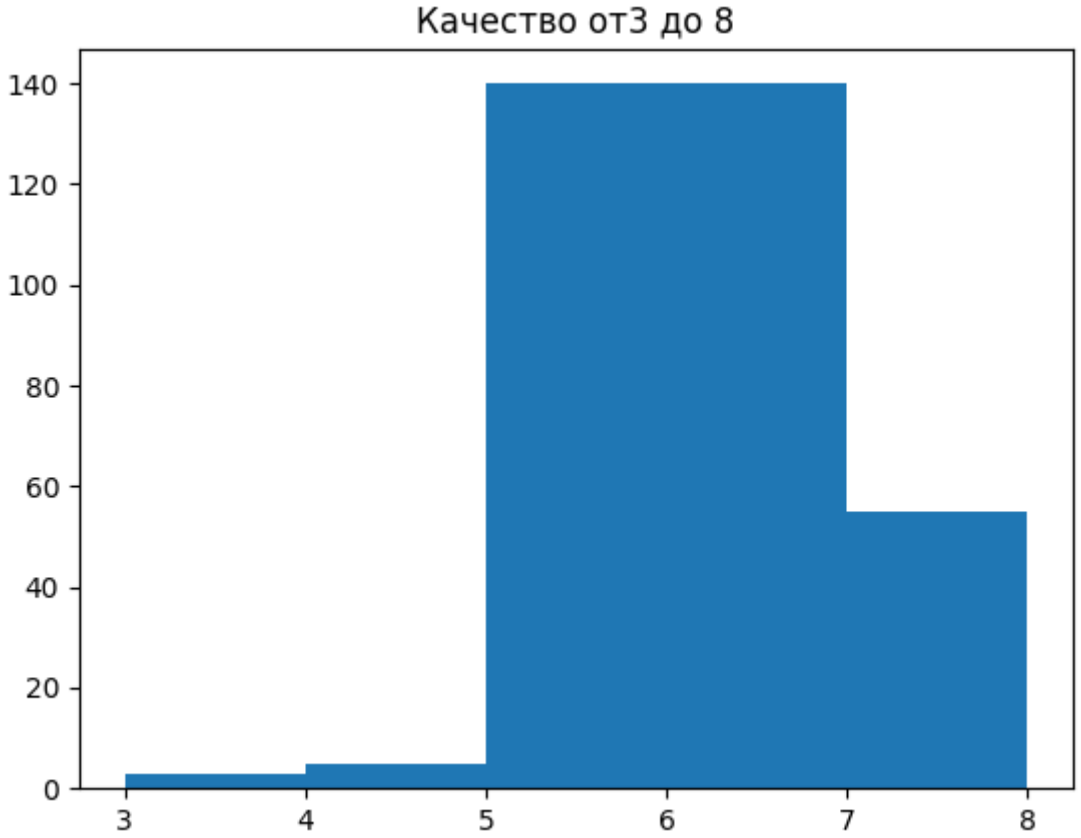
Массив тестовых значений



Массив прогнозируемых значений

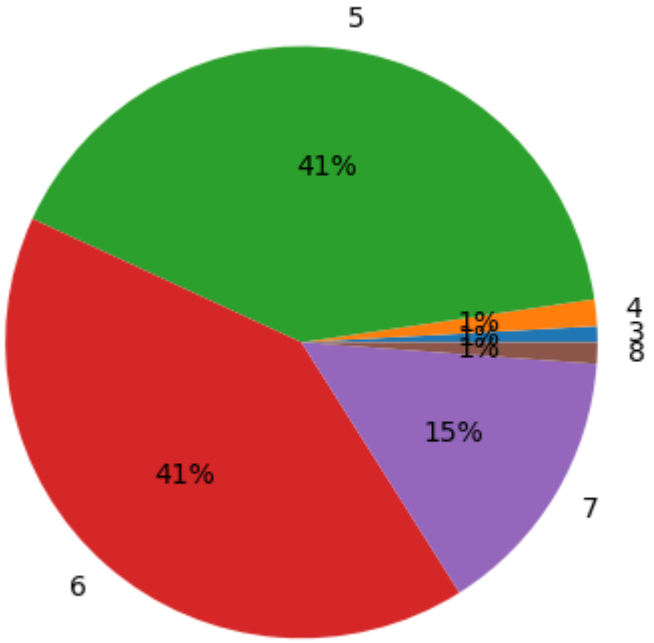




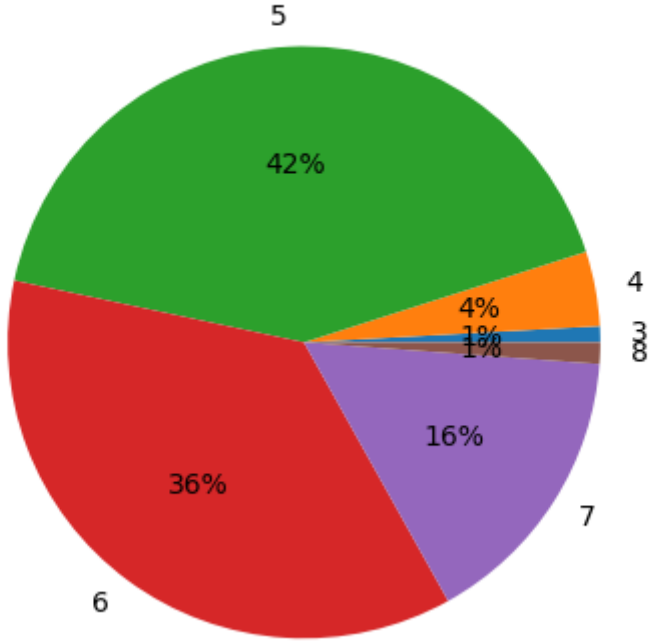


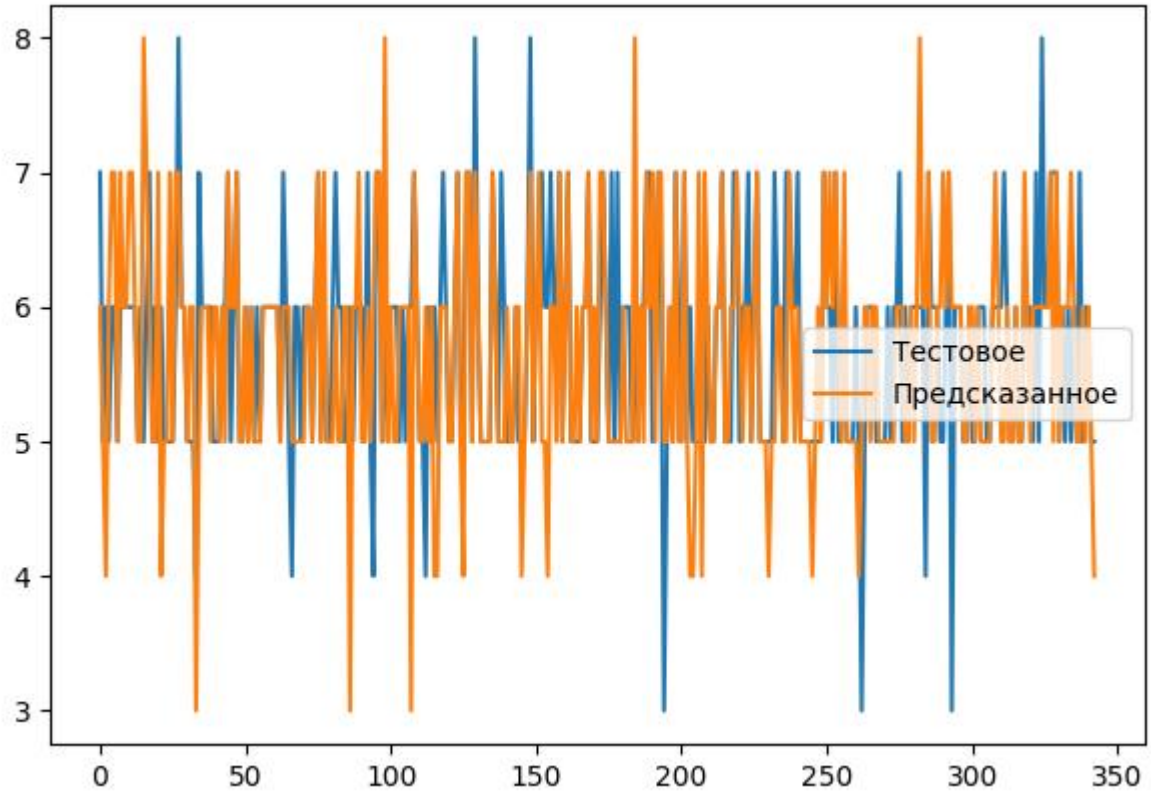
8. Провести предобработку данных.
9. Провести обучение и оценку моделей на очищенных данных.

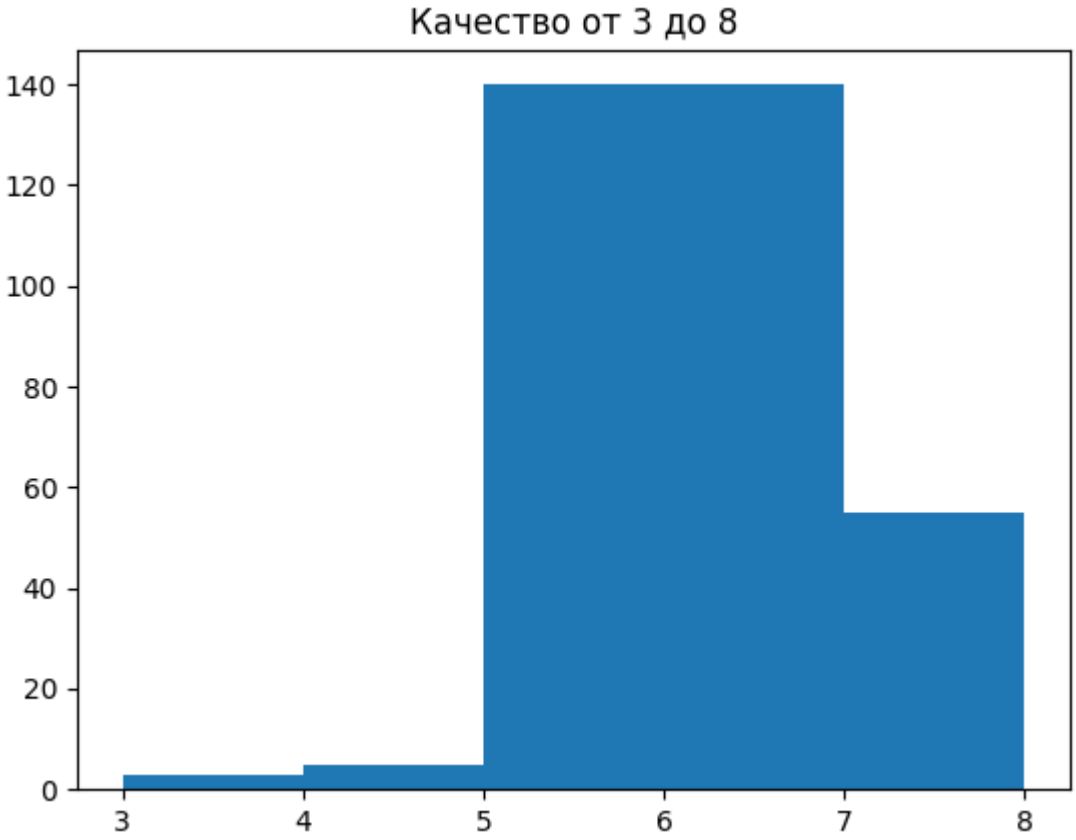
Массив тестовых значений



Массив прогнозируемых значений

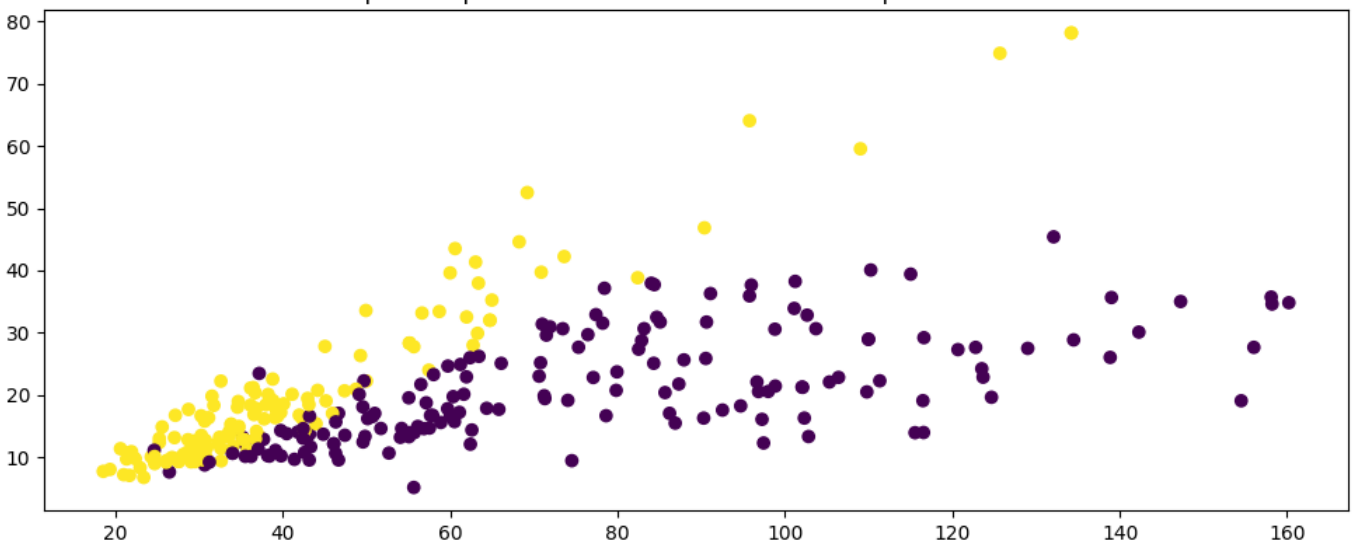




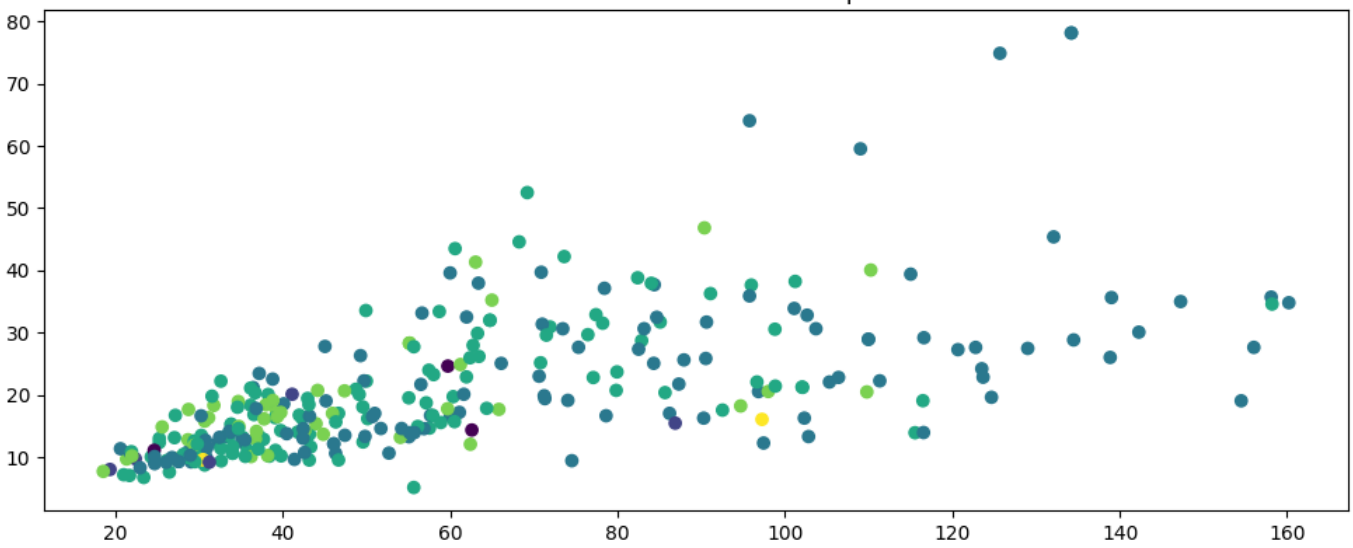


10. Проанализировать результаты.
11. Результаты анализа представить в табличной и графической форме.

Прогнозируемая зависимость качества красного вина



Фактическая зависимость качества красного вина



```
D:\lab-4\venv\Scripts\python.exe D:\lab-4\main.py
Просмотр пяти рядов
    fixed acidity  volatile acidity  citric acid  ...  alcohol  quality  Id
0             7.4              0.70         0.00  ...      9.4        5   0
1             7.8              0.88         0.00  ...      9.8        5   1
2             7.8              0.76         0.04  ...      9.8        5   2
3            11.2              0.28         0.56  ...      9.8        6   3
4             7.4              0.70         0.00  ...      9.4        5   4

[5 rows x 13 columns]
Формирование набора данных :  (1143, 13)
Количество отсутствующих значений во всем наборе данных
 fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
Id                      0
dtype: int64
Статистика
        fixed acidity  volatile acidity  citric acid  ...  alcohol  quality       Id
count         1143.00           1143.00      1143.00  ...  1143.00  1143.00  1143.00
mean             8.31              0.53         0.27  ...    10.44     5.66   804.97
std              1.75              0.18         0.20  ...     1.08     0.81   464.00
min              4.60              0.12         0.00  ...     8.40     3.00     0.00
25%              7.10              0.39         0.09  ...     9.50     5.00   411.00
50%              7.90              0.52         0.25  ...    10.20     6.00   794.00
75%              9.10              0.64         0.42  ...    11.10     6.00  1209.50
max             15.90              1.58         1.00  ...    14.90     8.00  1597.00

[8 rows x 13 columns]

Качество ценности :  [5 6 7 4 8 3]
Группировка по качеству
          fixed acidity  volatile acidity  citric acid  ...    pH  sulphates  alcohol
quality                                                 ...
3                 8.45              0.90         0.21  ...  3.36       0.55     9.69
4                 7.81              0.70         0.17  ...  3.39       0.64    10.26
5                 8.16              0.59         0.24  ...  3.30       0.61     9.90
6                 8.32              0.50         0.26  ...  3.32       0.68    10.66
7                 8.85              0.39         0.39  ...  3.29       0.74    11.48
8                 8.81              0.41         0.43  ...  3.24       0.77    11.94

[6 rows x 11 columns]
```

```
Отчет, показывающий основные метрики классификации
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00         5
           5       0.51      0.69      0.58       140
           6       0.47      0.45      0.46       140
           7       0.40      0.12      0.18        51
           8       0.00      0.00      0.00         4

    accuracy                           0.48       343
   macro avg       0.23      0.21      0.20       343
weighted avg       0.46      0.48      0.45       343

Матрица неточностей для оценки точности классификации [[ 0  0  2  1  0  0]
 [ 0  0  4  1  0  0]
 [ 0  1 97 42  0  0]
 [ 0  0 70 63  7  0]
 [ 0  1 18 26  6  0]
 [ 0  0  1  1  2  0]]
Оценка модели KNN: оценка точности  0.4839650145772595
 Отчет, показывающий основные метрики классификации               precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00         5
           5       0.64      0.66      0.65       140
           6       0.61      0.54      0.57       140
           7       0.52      0.55      0.53        51
           8       0.00      0.00      0.00         4

    accuracy                           0.57       343
   macro avg       0.29      0.29      0.29       343
weighted avg       0.59      0.57      0.58       343

 Матрица неточностей для оценки точности классификации [[ 0  0  2  1  0  0]
  [ 1  0  1  3  0  0]
  [ 0  9 92 31  7  1]
  [ 2  5 38 76 16  3]
  [ 0  0 10 13 28  0]
  [ 0  0  0  1  3  0]]
```

```
Основные метрики классификации
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00         5
           5       0.49      0.64      0.55       140
           6       0.42      0.48      0.45       140
           7       0.00      0.00      0.00        51
           8       0.00      0.00      0.00         4

    accuracy                           0.45       343
   macro avg       0.15      0.19      0.17       343
weighted avg       0.37      0.45      0.41       343

Матрица ошибок для оценки точности [[ 0  0  3  0  0  0]
 [ 0  0  2  3  0  0]
 [ 0  0 89 51  0  0]
 [ 0  0 73 67  0  0]
 [ 0  0 14 37  0  0]
 [ 0  0  1  3  0  0]]
Счет X-train с Y-train:  0.51
Счет X-test  с Y-test  :  0.45481049562682213
Точность метода логистической регрессии  0.45481049562682213

Размер массивов
X train :  (800, 11)
X test  :  (343, 11)
Y train :  (800,)
Y test  :  (343,)
```

12. Сформулировать выводы.

В ходе выполнения лабораторной работы, была создана программа с 3 классификаторами для машинного обучения. Была осуществлена отладка и работа, программы, зафиксированы показания, выводы, итоги, результаты. Было попунктно выполнено задание преподавателя и оформлено итоговый отчет. Все данные, файлы, исходники выгружены в директорию студента.