|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное автономное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | ГОЛОВНОЙ УЧЕБНО-ИССЛЕДОВАТЕЛЬСКИЙ И МЕТОДИЧЕСКИЙ ЦЕНТР |
| ПРОФЕССИОНАЛЬНОЙ РЕАБИЛИТАЦИИ ЛИЦ С ОГРАНИЧЕННЫМИ |
| ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ |
| КАФЕДРА | СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

|  |
| --- |
| ***Решение задачи машинного обучения*** |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ5Ц-84Б |  |  |  | К.Р. Падалко |
|  | (группа) |  | (подпись, дата) |  | (И.О. Фамилия) |
| Руководитель НИР |  |  |  |  | Ю.Е. Гапанюк |
|  |  |  | (подпись, дата) |  | (И.О. Фамилия) |
|  |  |  |  |  |  |

*2025 г.*

**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)**

|  |  |  |
| --- | --- | --- |
| Утверждаю | | |
|  | | |
| Заведующий кафедрой | | ИУ5 |
|  | | (индекс) |
|  | В.И. Терехов | |
|  | (И.О. Фамилия) | |
| (подпись) |  | |
|  | (дата) | |

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| по теме | | Решение задачи машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | ИУ5Ц-84Б | | | | Падалко Ксения Романовна | | | | | | | |
|  | | | | (Фамилия, имя, отчество) | | | | | | | | | | | |
| Направленность НИР (учебная, исследовательская, практическая, производственная, др.) | | | | | | | | | | | | | | | |
| ИССЛЕДОВАТЕЛЬСКАЯ | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| Источник тематики (кафедра, предприятие, НИР) | | | | | | | | | | КАФЕДРА | | | | | |
| График выполнения НИР: | | | | | |  | | | | | | | | | |
| 25% к | 3 | | нед., 50% к | | | 9 | нед., 75% к | | 12 | | | нед., 75% к | 15 | нед |
| ***Техническое задание:*** | | | | |  | | | | | | | | | | |
| Решение задачи машинного обучения. Результатом проекта является отчет, содержащий описания моделей, тексты программ и результаты экспериментов. | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***Оформление научно-исследовательской работы:*** | | | | | | | | | | |  | | | | |
| Расчетно-пояснительная записка на \_42\_\_\_\_ листах формата А4. | | | | | | | | | | | | | | | |
| Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.) | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |

Дата выдачи задания « 07 » \_февраля\_\_\_ 2025 г.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Руководитель НИР** |  |  |  | Ю.Е. Гапанюк |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |
| **Студент** |  |  |  | К.Р. Падалко |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**СОДЕРЖАНИЕ**

[Введение 4](#_Toc195769488)

[Основная часть 6](#_Toc195769489)

[1. Поиск и выбор набора данных для построения моделей машинного обучения. 6](#_Toc195769490)

[2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных. 7](#_Toc195769491)

[3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей. 18](#_Toc195769492)

[4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. 21](#_Toc195769493)

[5. Выбор метрик для последующей оценки качества моделей. 24](#_Toc195769494)

[6. Выбор наиболее подходящих моделей для решения задачи классификации. 26](#_Toc195769495)

[7. Формирование обучающей и тестовой выборок на основе исходного набора данных. 26](#_Toc195769496)

[8. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки. 27](#_Toc195769497)

[9. Подбор гиперпараметров для выбранных моделей. 31](#_Toc195769498)

[10. Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей. 35](#_Toc195769499)

[11. Формирование выводов о качестве построенных моделей на основе выбранных метрик. 36](#_Toc195769500)

[12. Вывод 39](#_Toc195769501)

[Заключение 41](#_Toc195769502)

[Список использованных источников информации 42](#_Toc195769503)

## Введение

Данное исследование основано на анализе набора данных о качестве вина, который представляет собой ценный ресурс для изучения факторов, влияющих на органолептические характеристики вин. Этот набор данных содержит информацию о различных химических и физико-химических свойствах вина, что позволяет глубже понять, как эти факторы влияют на общее восприятие качества продукта.

Целевым признаком для данного исследования выбрана переменная "quality" — оценка качества вина, основанная на дегустации. Этот выбор обусловлен важностью понимания факторов, влияющих на восприятие качества вина, и их взаимодействия с конкретными химическими характеристиками.

В рамках исследования мы сосредоточимся на анализе структуры данных и основных характеристик набора данных:

* fixed acidity — фиксированная кислотность;
* volatile acidity — летучая кислотность;
* citric acid — лимонная кислота;
* residual sugar — остаточный сахар;
* chlorides — хлористые соединения;
* free sulfur dioxide — свободный двуокись серы;
* total sulfur dioxide — общий диоксид серы;
* density — плотность;
* pH — уровень pH;
* sulphates — сульфаты;
* alcohol — содержание алкоголя;
* quality — оценка качества (от 0 до 10).

Анализ этих характеристик позволит выявить возможные корреляции между химическими параметрами и качеством вина, а также определить наиболее значимые факторы, оказывающие влияние на его восприятие.

Таким образом, данное исследование расширит наше понимание химических аспектов, определяющих качество вина, и предоставит основы для разработки рекомендаций по улучшению производства винодельческой продукции.

Поставим цели, которых мы хотим добиться, выполнив данную научно-исследовательскую работу.

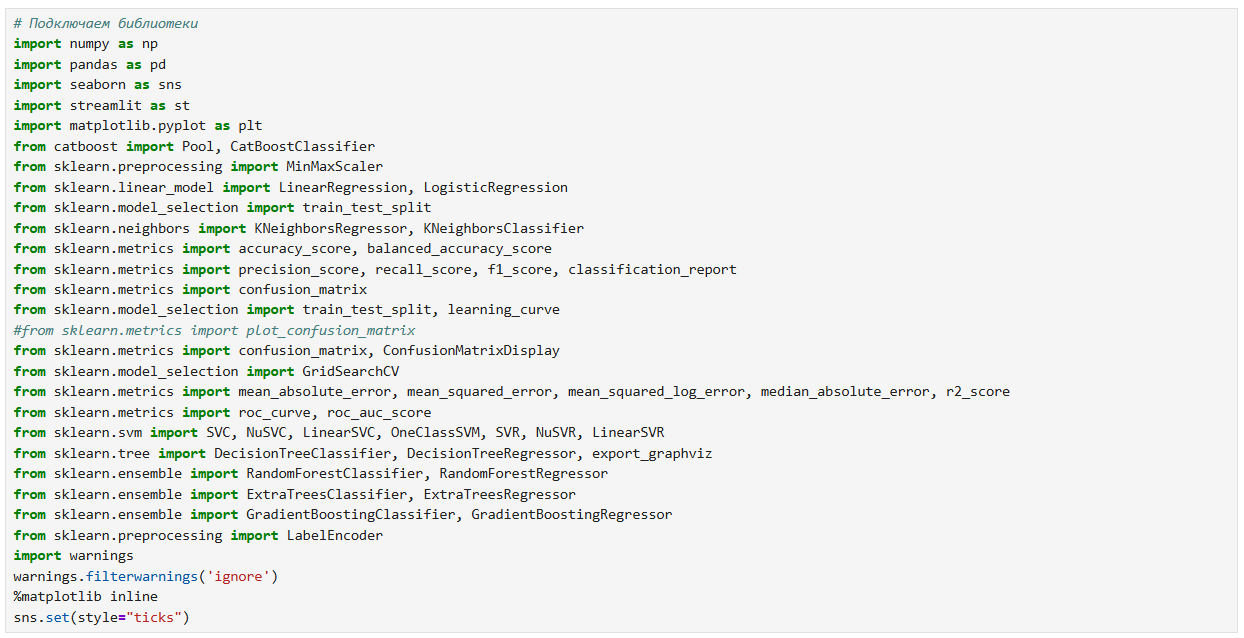
Цели:

1. Поиск и выбор набора данных для построения моделей машинного обучения.
2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.
3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
5. Выбор метрик для последующей оценки качества моделей.
6. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
7. Формирование обучающей и тестовой выборок на основе исходного набора данных.
8. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки.
9. Подбор гиперпараметров для выбранных моделей.
10. Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

## Основная часть

Ссылка на датасет: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

## Поиск и выбор набора данных для построения моделей машинного обучения.



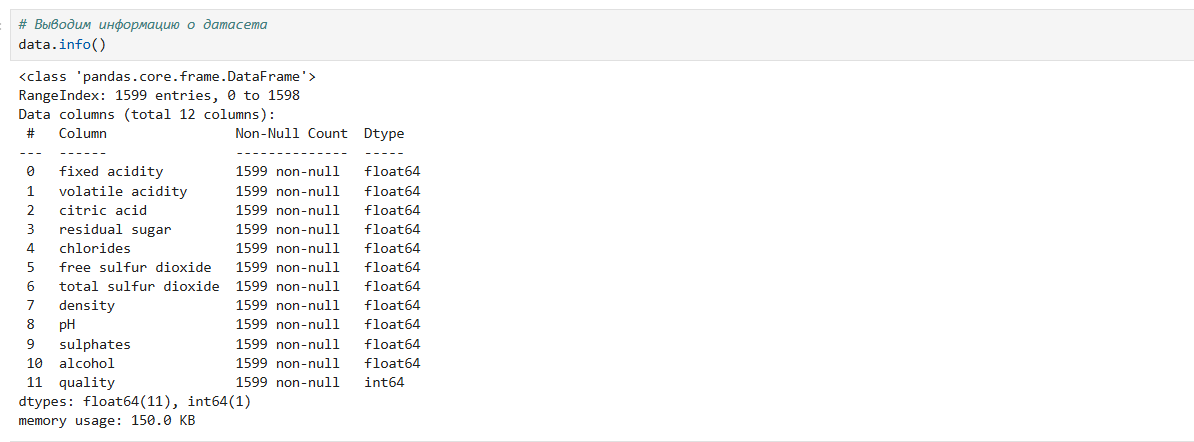
Здесь я переименовал названия каждого столбца, так как в источнике данных не было конкретных наименований для столбцов.



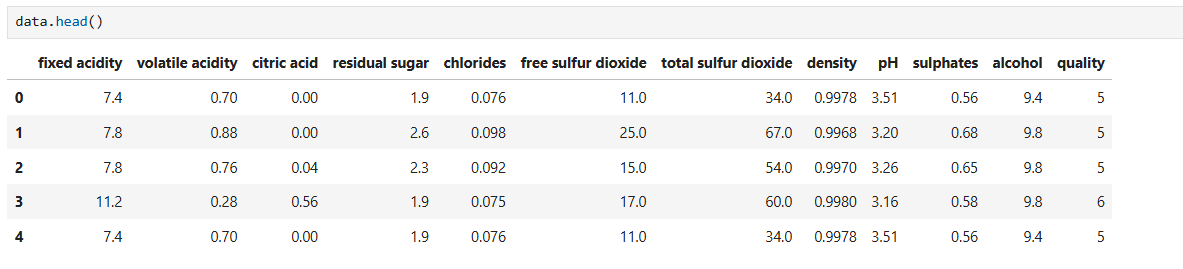
Файл успешно загружен, теперь рассматриваем следующее.

## Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

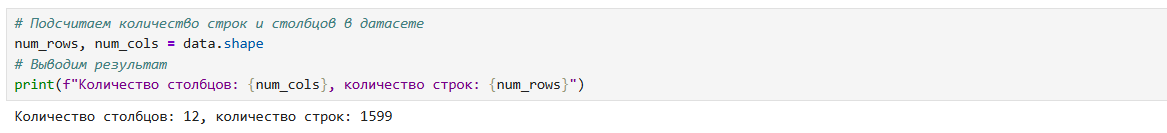
Выводим информацию о датасете.



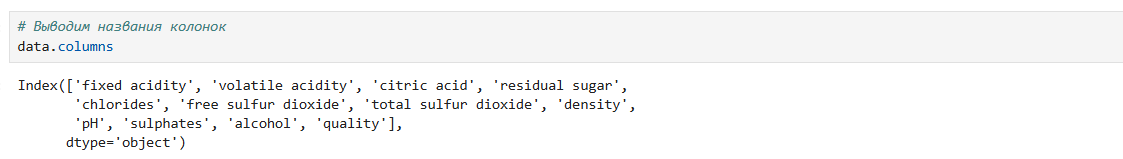
Выводим первые 5 строк датасета.



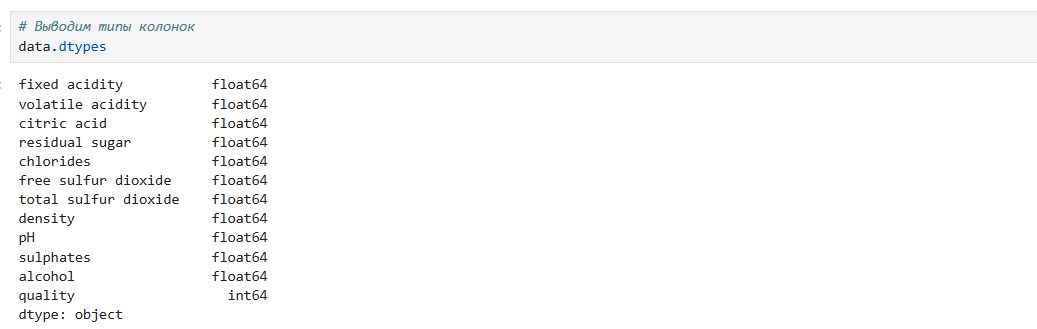
Подсчитаем количество строк и столбцов в датасете и выводим результат.



Выводим названия колонок.

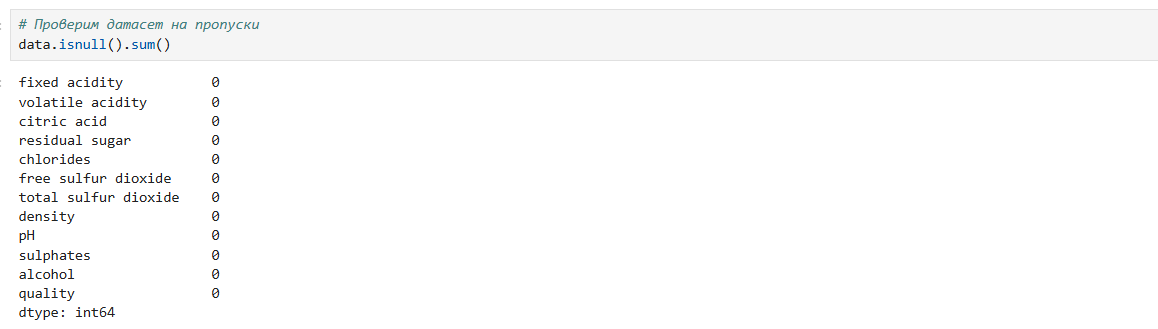


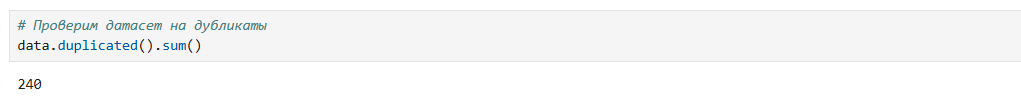
Выводим типы колонок.



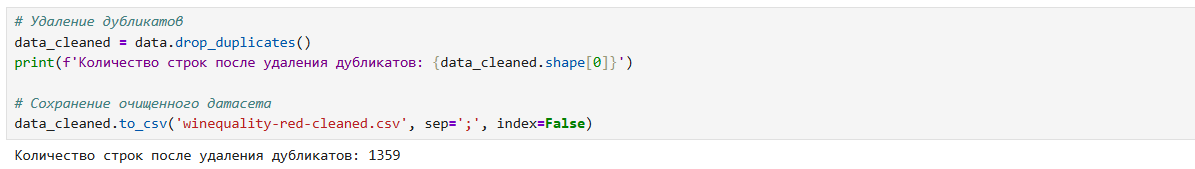
Практически в каждом колонке имеется тип float64, это значит, что все колонки имеют вещественные числа.

Проверим датасет на пропуски и на дубликаты.

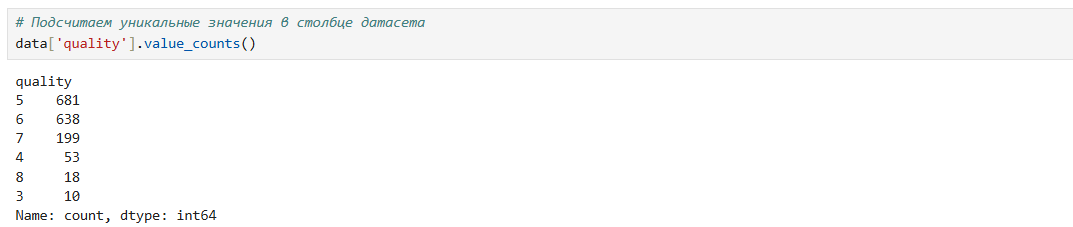




Пропусков нет, а дубликаты есть. Удаляем дубликаты.



Подсчитаем уникальные значения в столбце датасета.

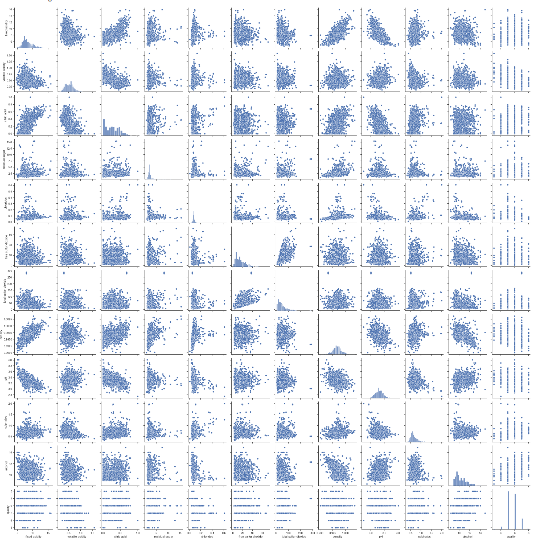


В столбце 'quality' значение '5' встречается 681 раз, а значение '6' - 638 раз, '7' - 199 раз, '4' - 53 раз, '8' - 18 раз, '3' - 10 раз.

Набор данных не содержит пропусков, категориальные признаки закодированы.

Построим графики рассеяния для всех возможных пар числовых столбцов.

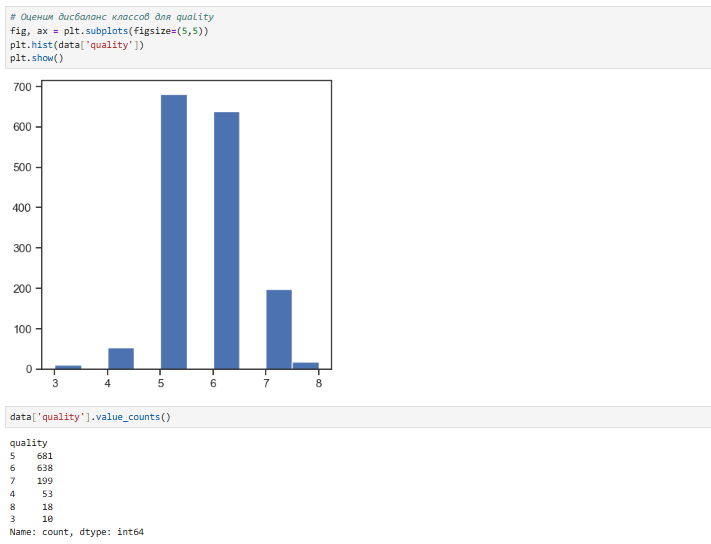




Построим графики рассеяния для всех возможных пар числовых столбцов quality.



Оценим дисбаланс классов для quality.



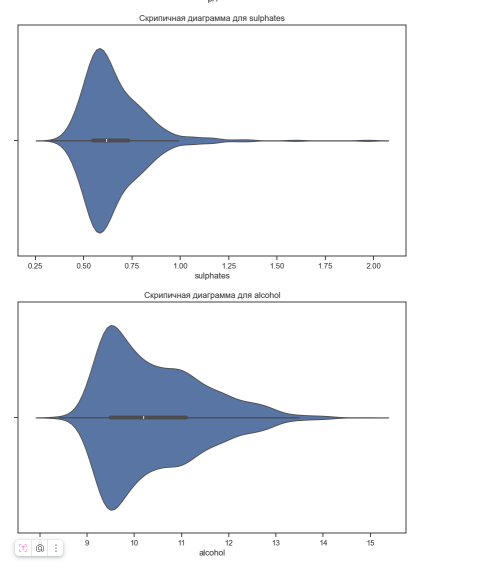
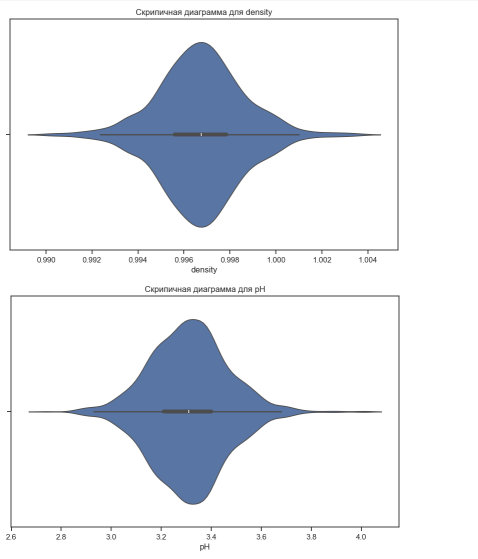
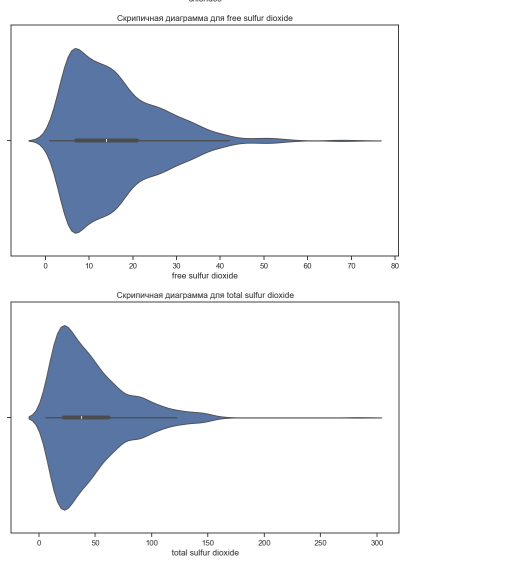
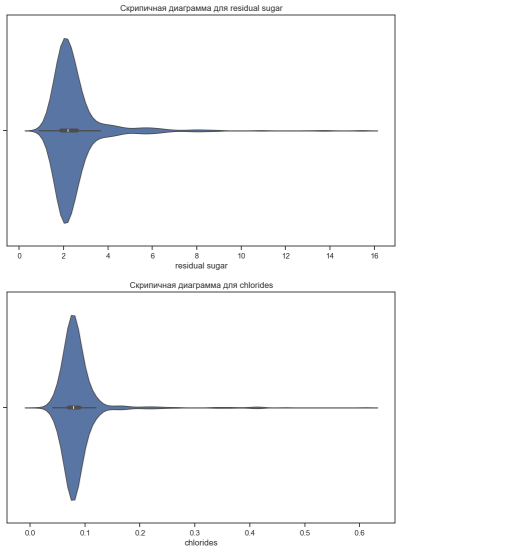
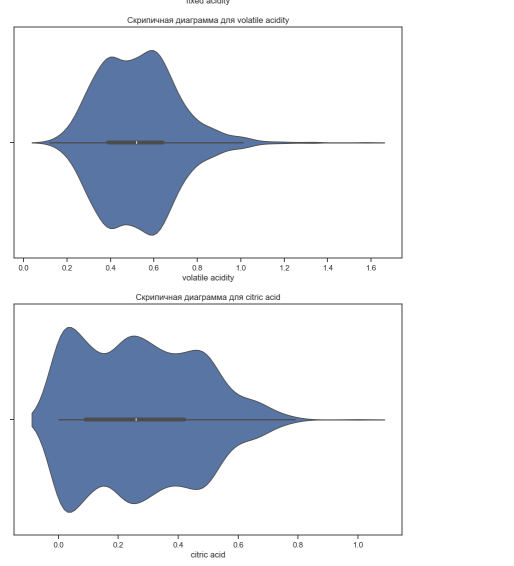
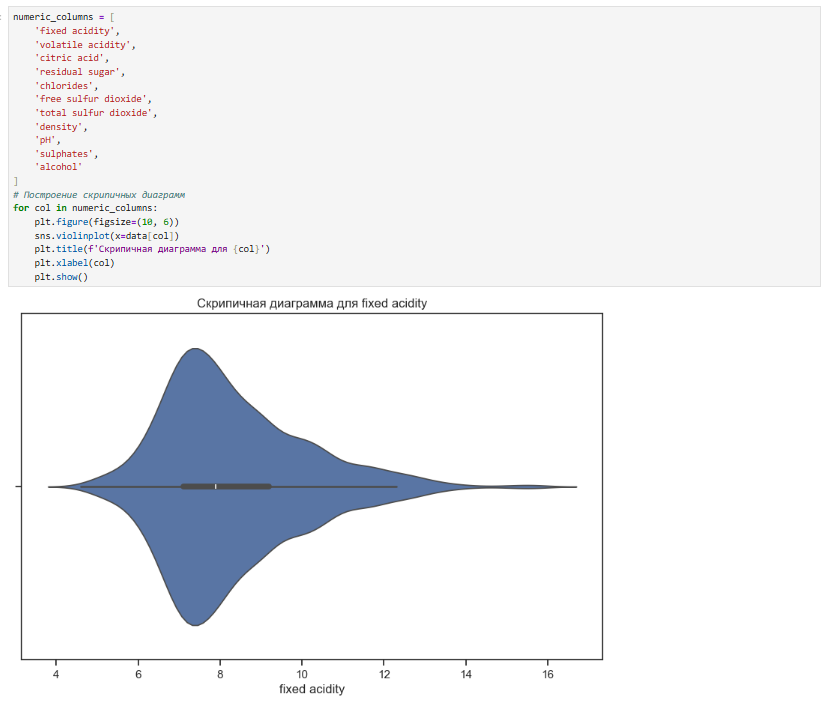
Подсчитаем дисбаланс классов.



В данном случае, качество 8 составляет примерно 42.59% от общего количества экземпляров, в то время как качество 7 составляет примерно 39.90%. Качество 6 составляет 12.45%, а более низкие качества (5, 4 и 3) составляют значительно меньшую долю — 3.31%, 1.13% и 0.63% соответственно.

Вывод подтверждает наличие дисбаланса качеств в данных. Преобладание качеств 8 и 7 над остальными может оказать влияние на способность модели классификации корректно идентифицировать все уровни качества. В частности, поскольку качества 8 и 7 составляют значительную часть выборки, модель может быть склонна к неправильному классифицированию экземпляров, относящихся к более низким качествам (5, 4 и 3), как к качествам 8 или 7. Это может привести к снижению общей эффективности модели в предсказании всех уровней качества вина.

Скрипичные диаграммы для числовых колонок.

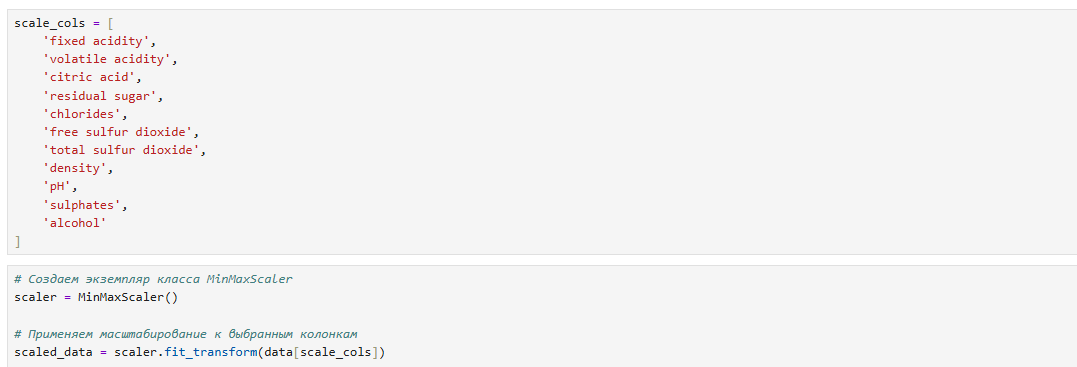


1. Различие в распределении: Если формы скрипичных диаграмм для разных классов значительно отличаются (например, одна скрипичная диаграмма имеет более широкий разброс значений или имеет более выраженные хвосты), это может указывать на различия в распределении признаков между классами.
2. Выбросы и экстремальные значения: Различия в форме скрипичных диаграмм могут также указывать на наличие выбросов или экстремальных значений в одном из классов.
3. Потенциальная предсказательная сила признаков: Если форма скрипичной диаграммы для одного из классов более выражена или имеет более четкие пики, чем для другого класса, это может указывать на то, что эти признаки имеют более высокую предсказательную силу для одного из классов.
4. Важность признаков: Различия в форме скрипичных диаграмм могут помочь определить, какие признаки могут быть более важными для разделения классов.

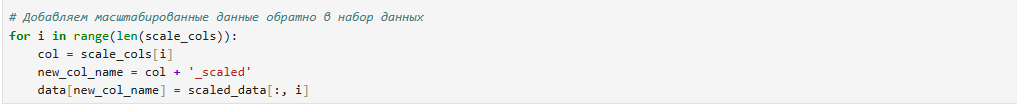
## Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.

Для построения моделей будем использовать все признаки. Категориальные признаки закодированы, поэтому выполним масштабирование данных.

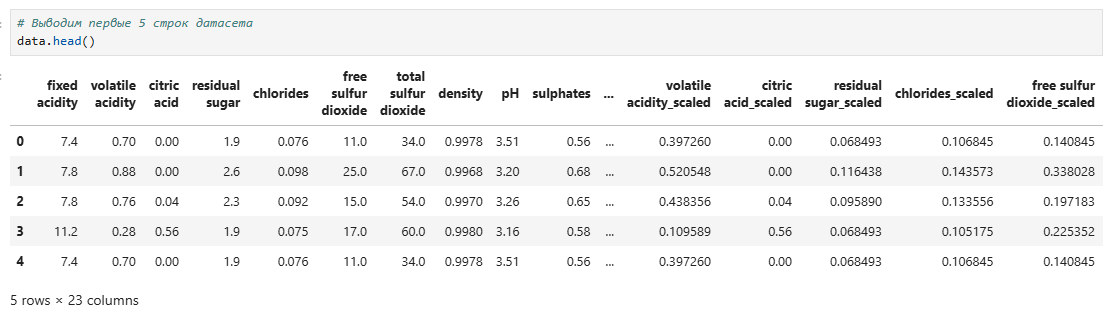
Числовые колонки для масштабирования данных.



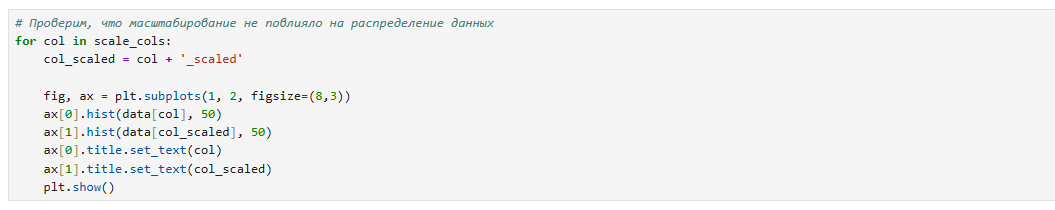
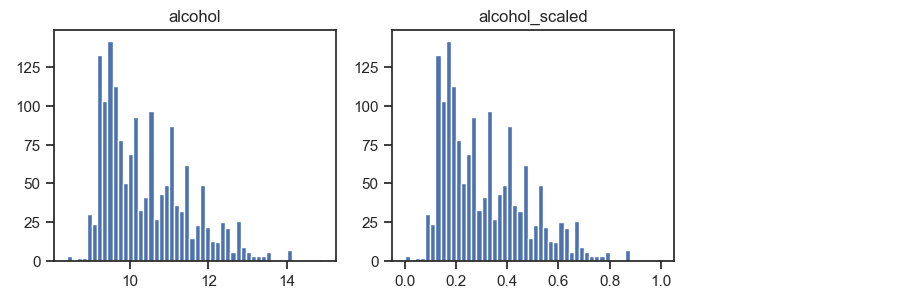
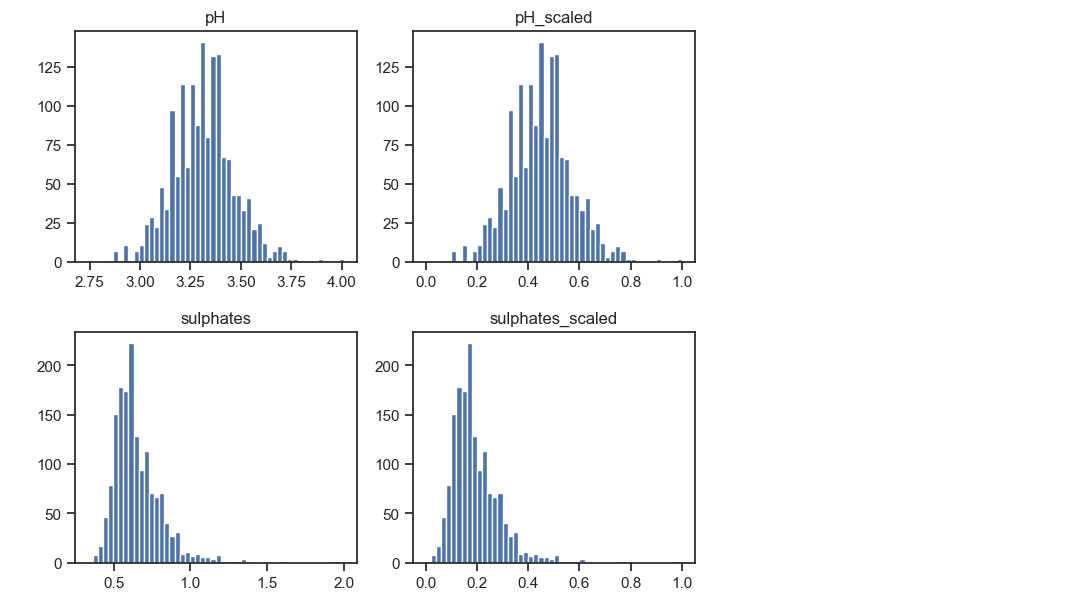
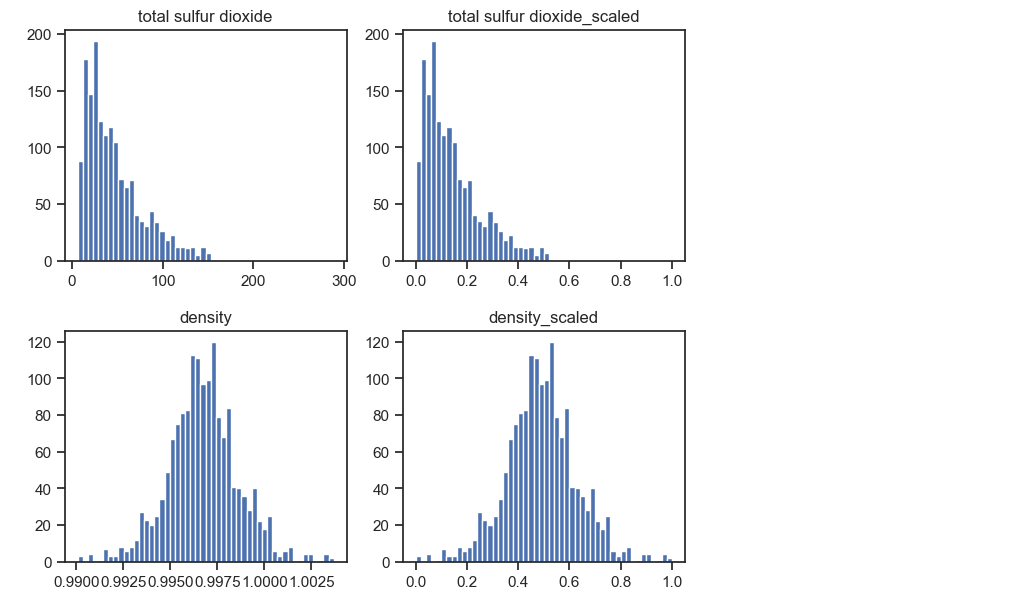
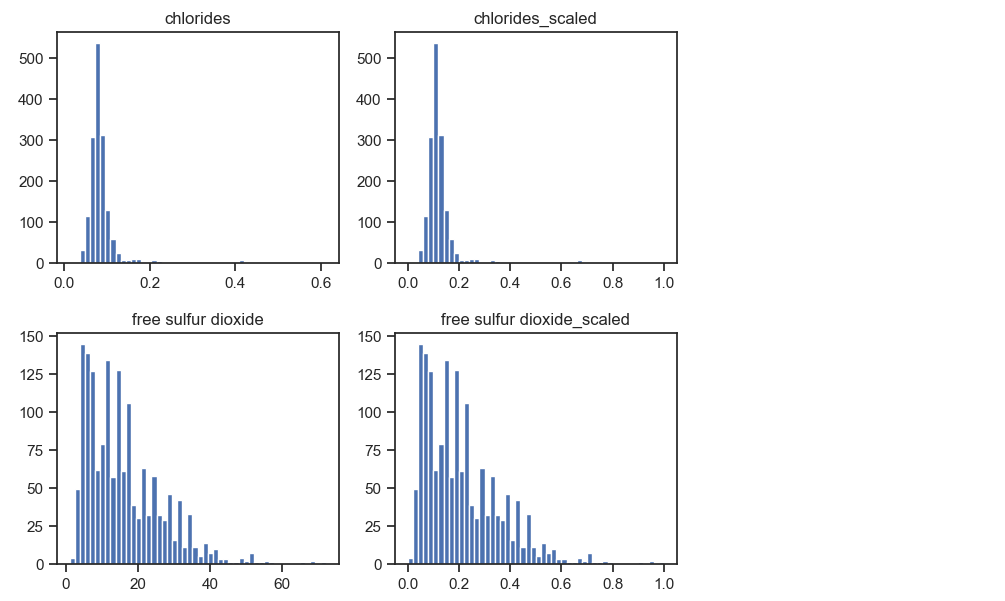
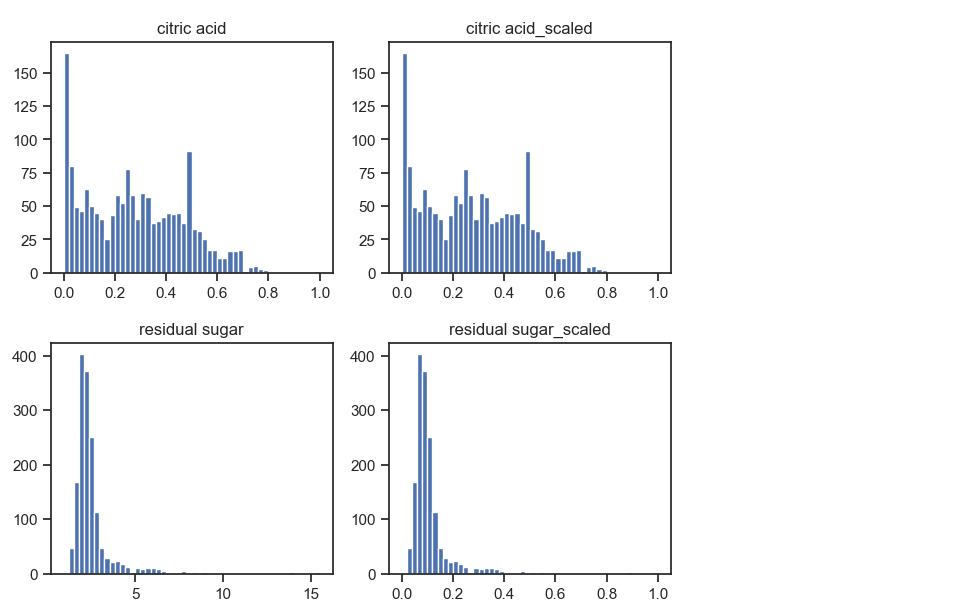
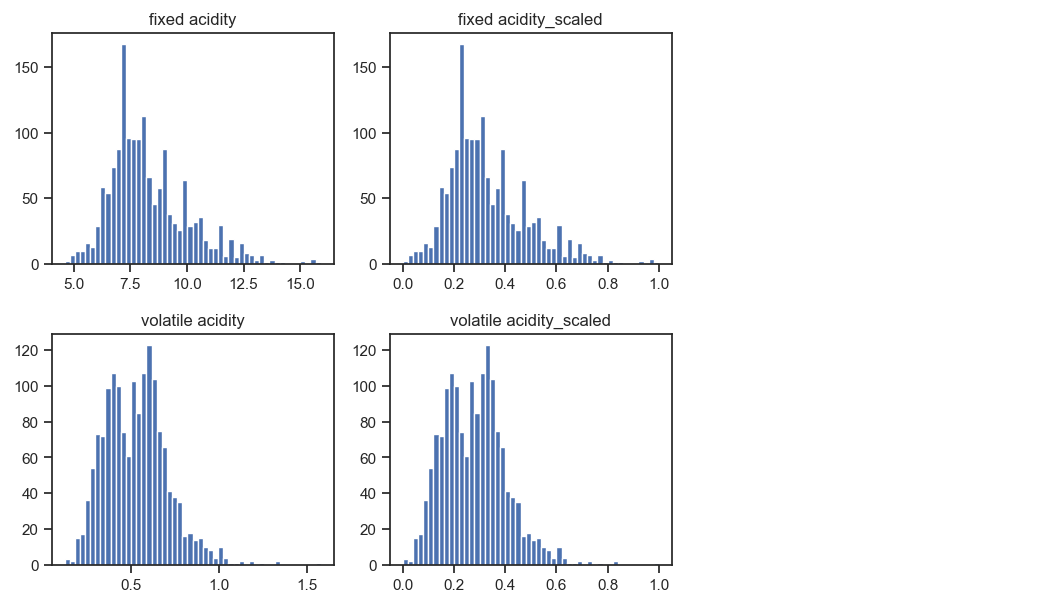
Добавим масштабированные данные в набор данных.



Выводим первые 5 строк датасета.



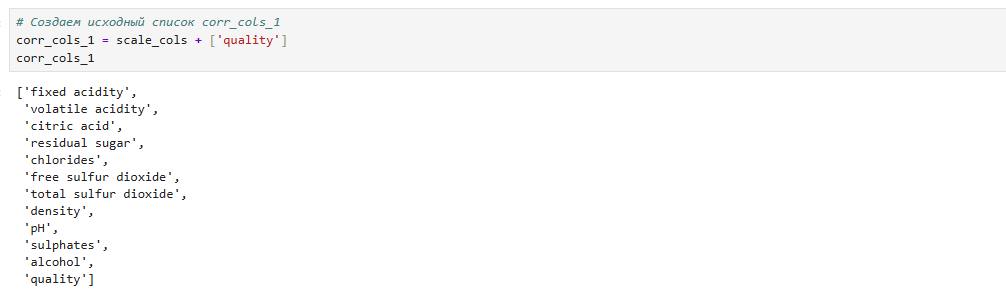
Проверим, что масштабирование не повлияло на распределение данных.

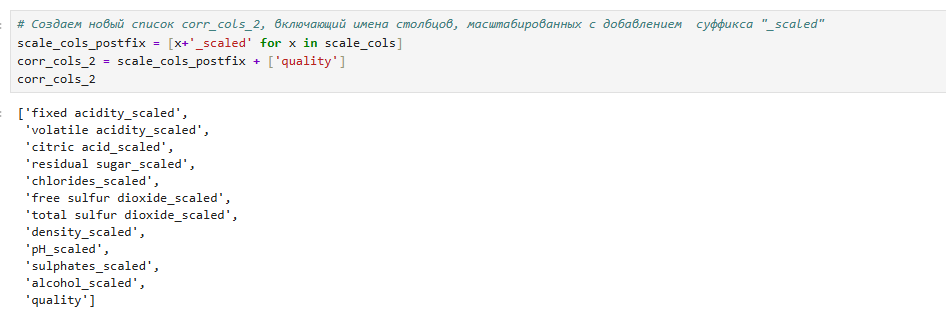
Все диаграммы имеют нормальный вид распределения.

## Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.

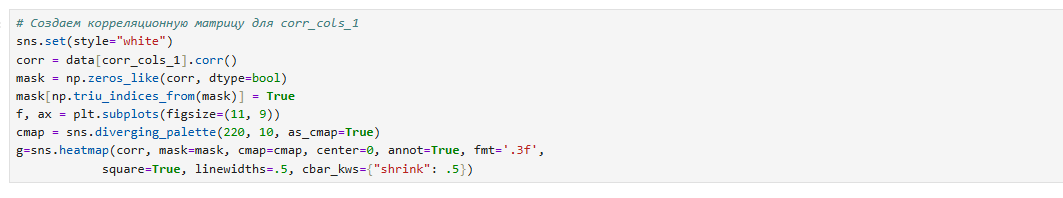
Создаем исходный список corr\_cols\_1.



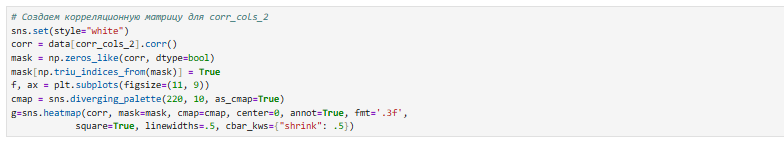
Создаем новый список corr\_cols\_2, включающий имена столбцов, масштабированных с добавлением суффикса "\_scaled".



Создаем корреляционную матрицу для corr\_cols\_1.



Создаем корреляционную матрицу для corr\_cols\_2.



На основе двух корреляционных матриц для сравнения сделаем следующие выводы:

Корреляционные матрицы для исходных и масштабированных данных совпадают. Целевой признак классификации "quality" наиболее сильно коррелирует со следующими признаками:

* "alcohol" (0.476);
* "sulphates" (0.251);
* "citric acid" (0.226).

На основании корреляционной матрицы можно сделать вывод о том, что данные позволяют построить модель машинного обучения.

## Выбор метрик для последующей оценки качества моделей.

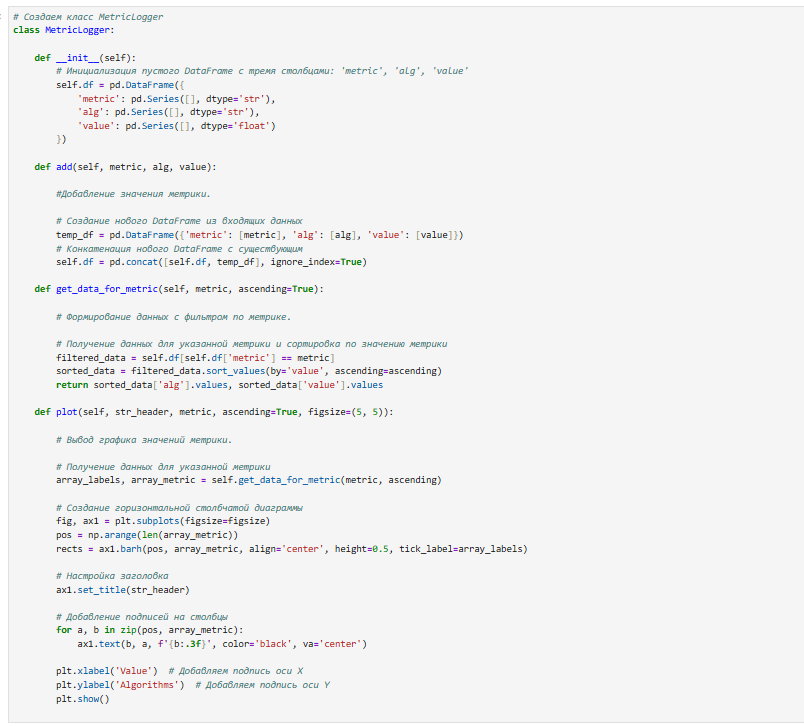
В качестве метрик для решения задачи классификации будем использовать:

* Precision - доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор определил как положительные. В контексте анализа качества вина это будет означать, сколько из предсказанных «хороших» вин действительно оказались хорошими.
* Recall - доля верно предсказанных классификатором положительных объектов, из всех действительно положительных объектов. Для нашей задачи это означает, насколько успешно модель находит все хорошие вина в датасете.
* F1- мера - для объединения precision и recall в единую метрику, обеспечивая баланс между ними.
* ROC AUC основана на вычислении следующих характеристик:
  + True Positive Rate, откладывается по оси ординат. Совпадает с recall.
  + False Positive Rate, откладывается по оси абсцисс. Показывает, какую долю из объектов отрицательного класса алгоритм предсказал неверно.

Отрисовка ROC-кривой.



Создаем класс MetricLogger. Это нужно для того, чтобы построить различные решения задач и оценивания качества моделей.



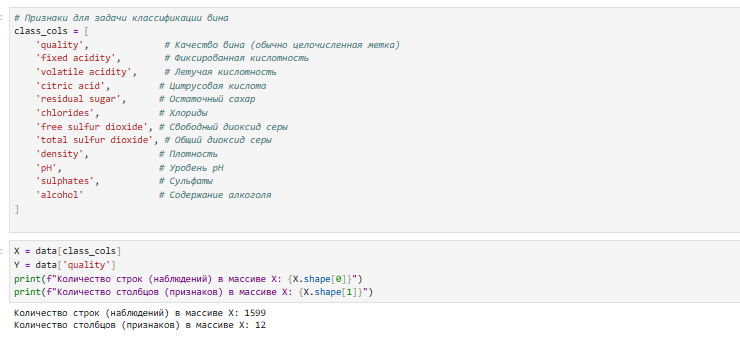
## Выбор наиболее подходящих моделей для решения задачи классификации.

Для задачи классификации будем использовать следующие модели:

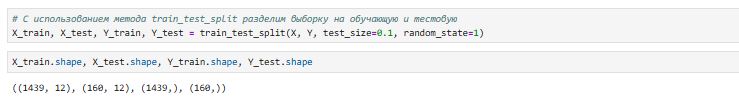
* Метод ближайших соседей
* Линейная/логистическая регрессия
* Метод опорных векторов
* Дерево решений
* Случайный лес
* Градиентный бустинг

## Формирование обучающей и тестовой выборок на основе исходного набора данных.

Признаки для задачи классификации.



Разделим выборку на обучающую и тестовую.



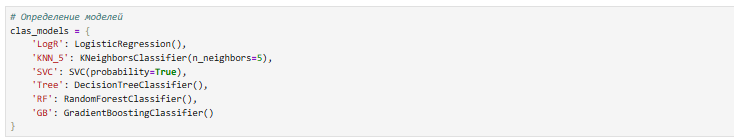
Из итогов ((1439, 12), (160, 12), (1439,), (160,)) можно сделать следующие выводы:

* X\_train содержит 1439 наблюдений и 12 признаков.
* X\_test содержит 160 наблюдений и 12 признаков.
* Y\_train содержит 1439 меток классов.
* Y\_test содержит 160 меток классов.

Это означает, что данные были успешно разделены на обучающую и тестовую выборки. Обучающая выборка используется для обучения модели, а тестовая выборка - для оценки ее качества на данных, которые модель ранее не видела.

## Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки.

Модели.



Сохранение метрик.



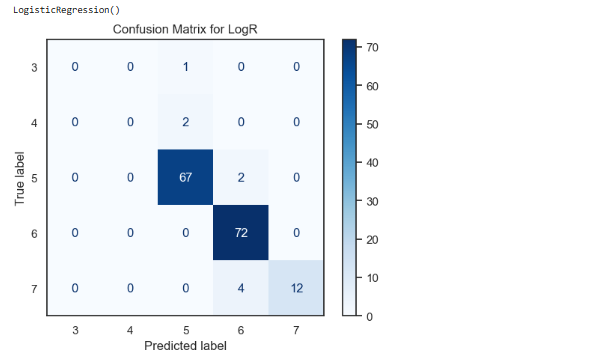
Обучающая модель.



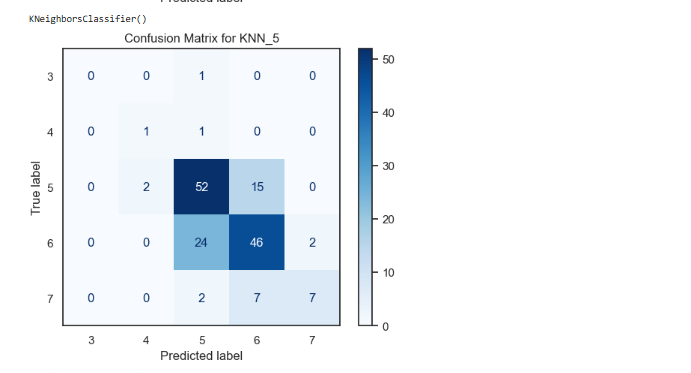
Выполним обучение каждой модели из словаря clas\_models, вычисляем и сохраняем метрики качества и результаты оценки производительности моделей.



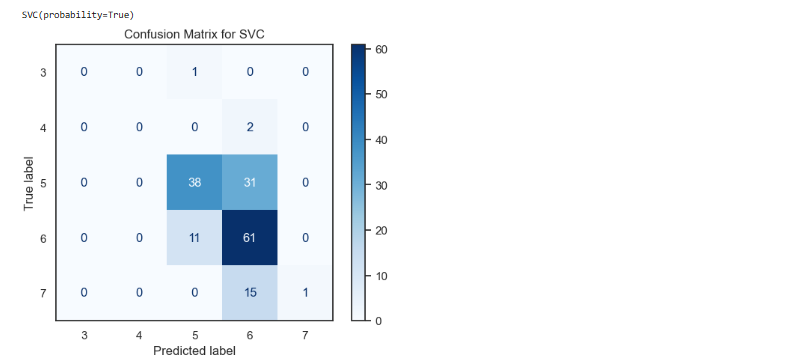
Метод ближайших соседей.



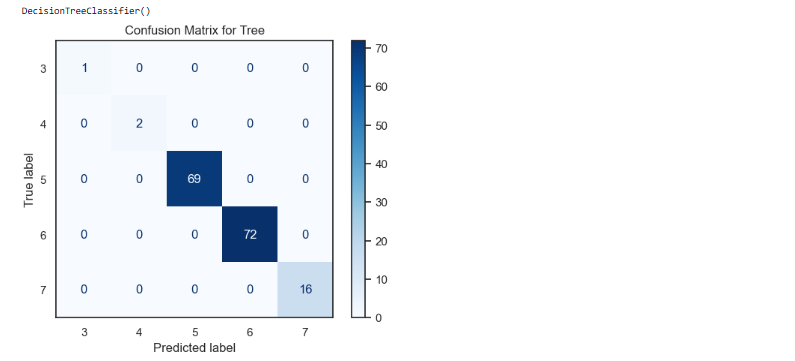
Логическая регрессия.



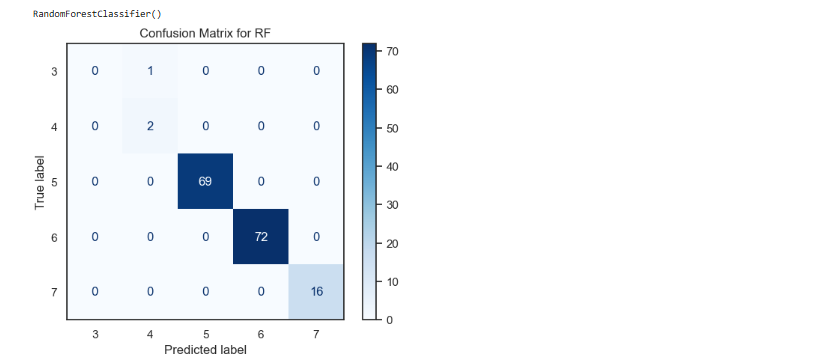
Метод опорных векторов.

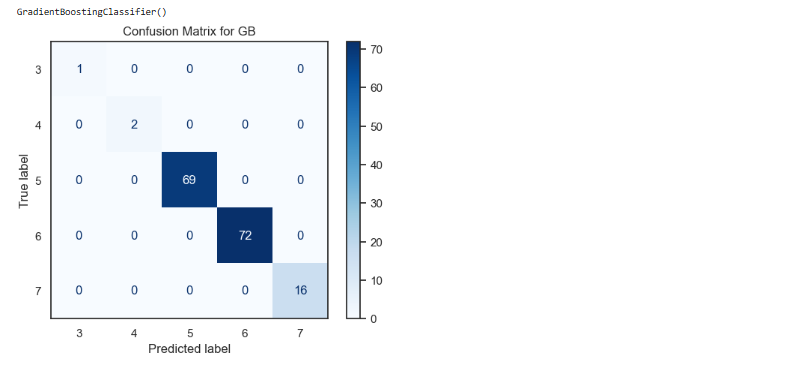


Дерево решений.



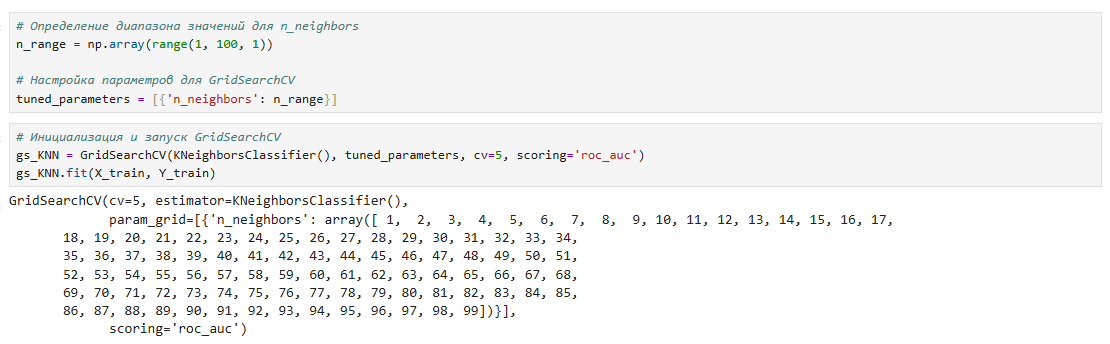
Случайный лес.

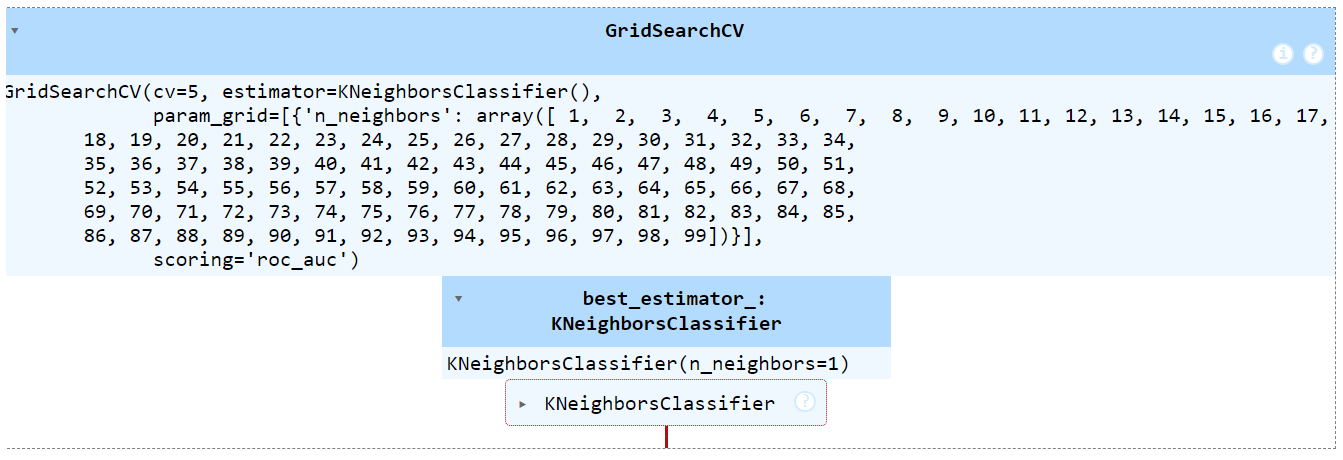


Градиентный бустинг. 

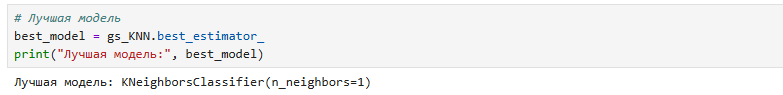
## Подбор гиперпараметров для выбранных моделей.

**Метод ближайших соседей.**

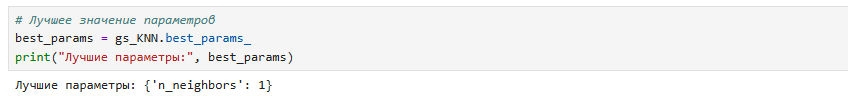




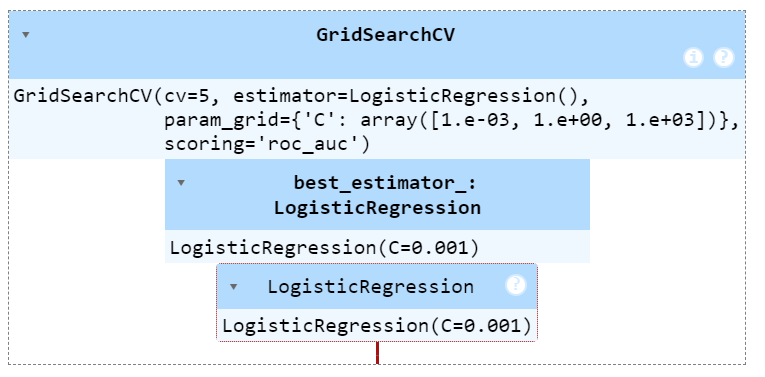
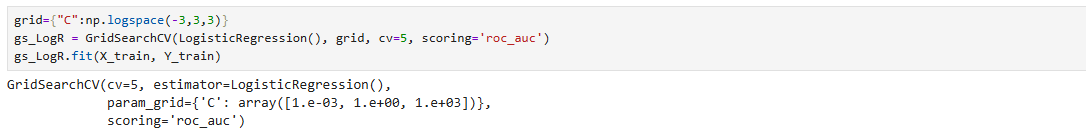
Лучшая модель.



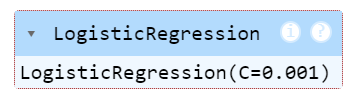
Лучшее значение параметров.



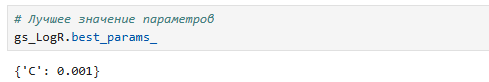
**Логистическая регрессия.**



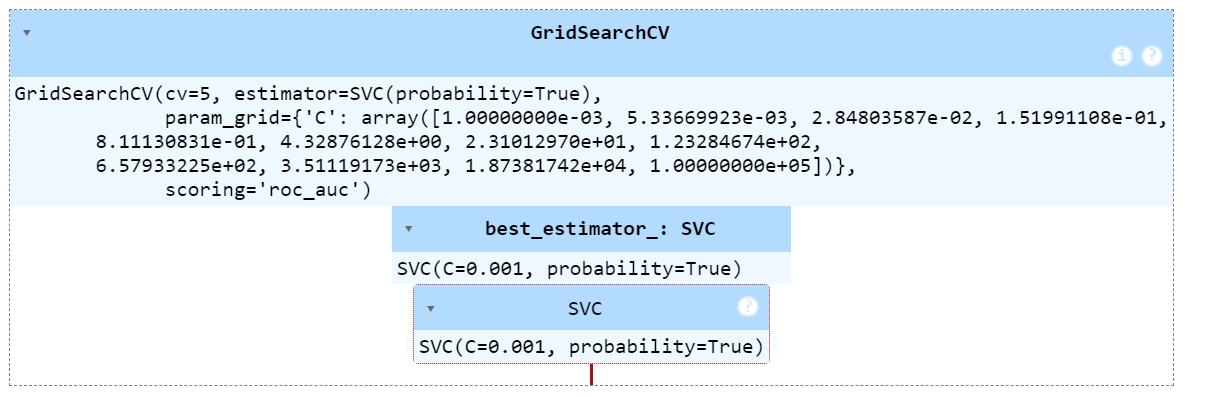
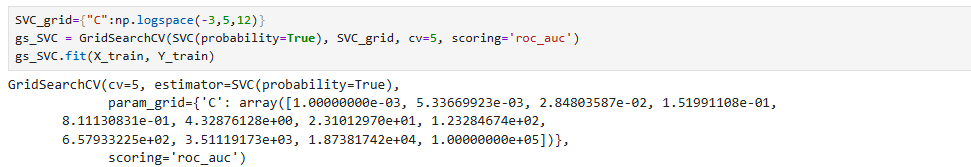
Лучшая модель.



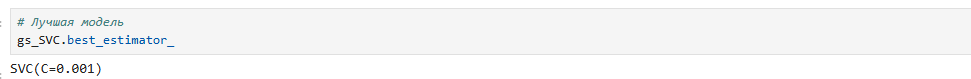
Лучшее значение параметров.



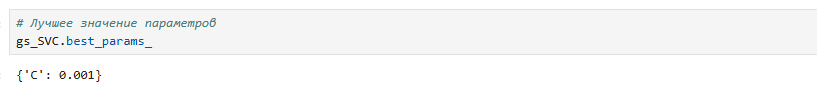
**Метод опорных векторов.**



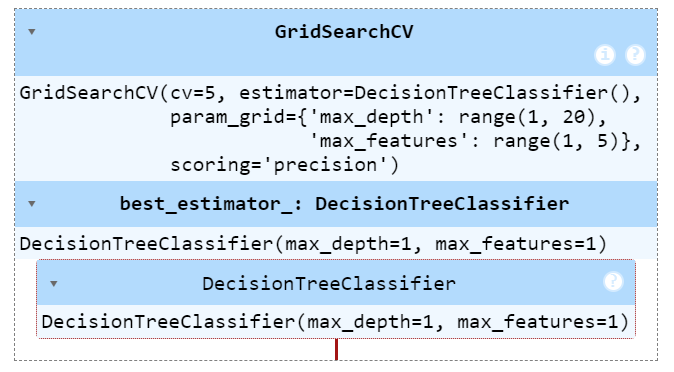
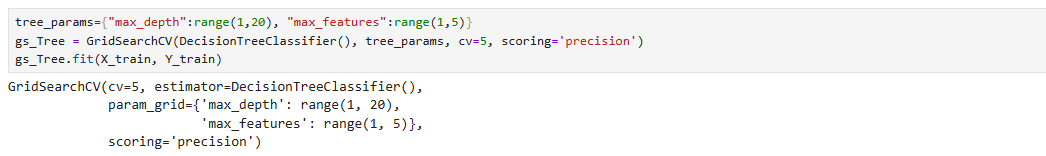
Лучшая модель.



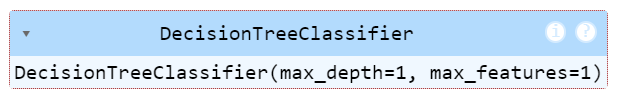
Лучшее значение параметров.



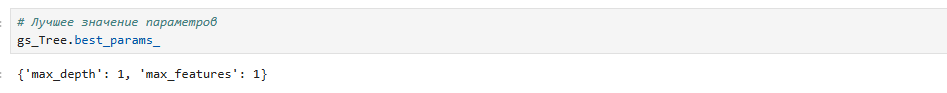
**Дерево решений.**



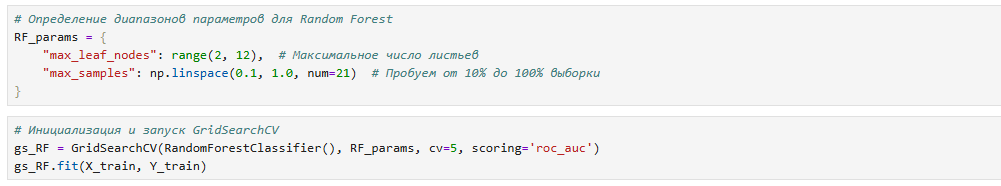
Лучшая модель.

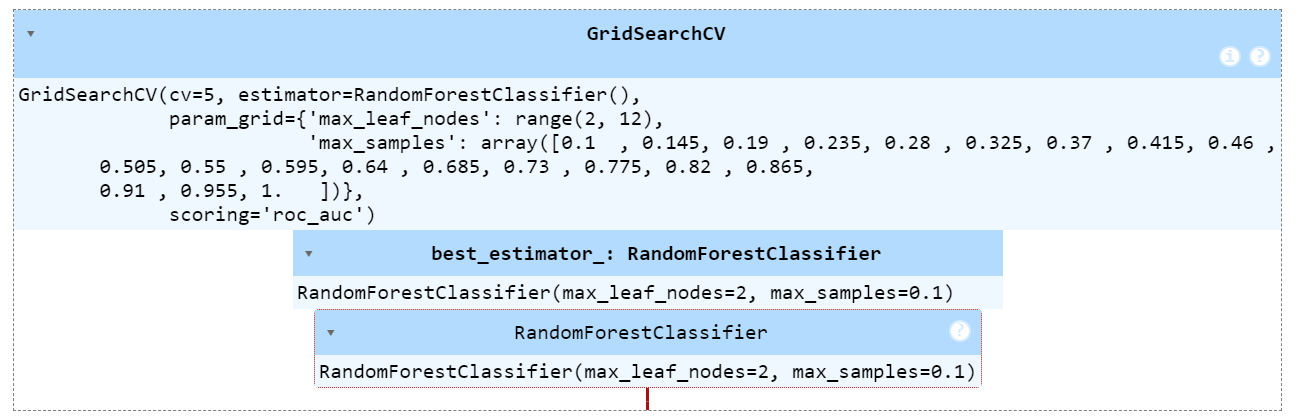


Лучшее значение параметров.

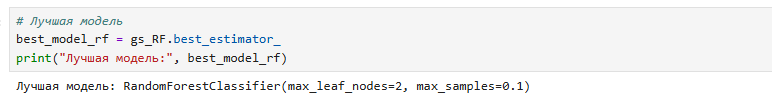


**Случайный лес.**





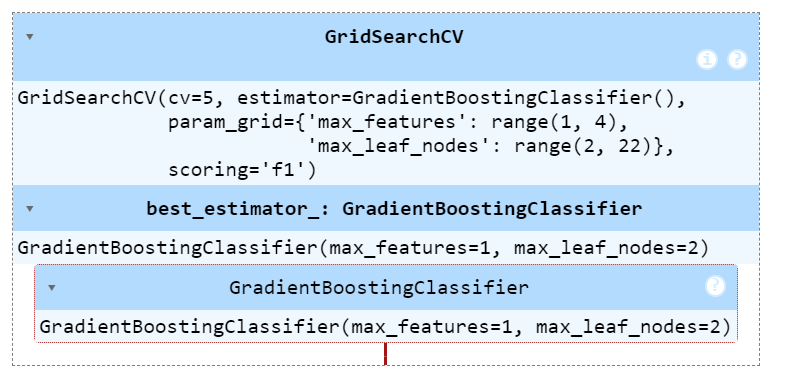
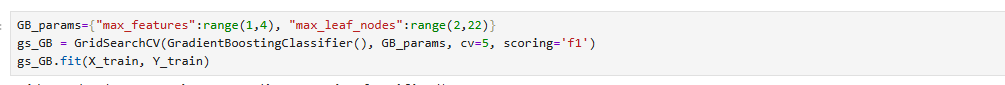
Лучшая модель.



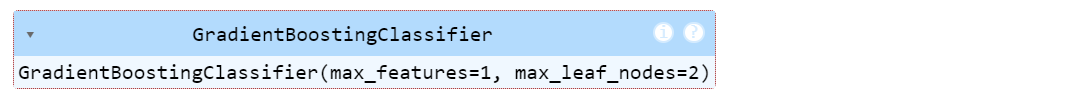
Лучшее значение параметров.



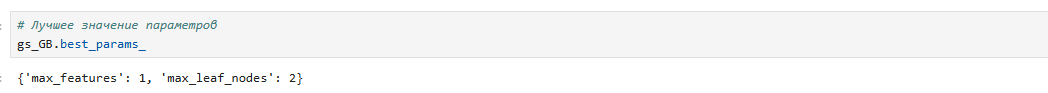
**Градиентный бустинг.**



Лучшая модель.



Лучшее значение параметров.

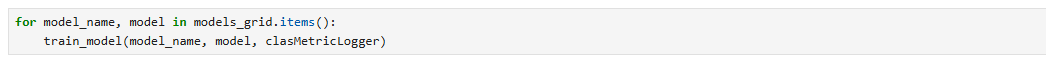


## Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.

Предположительно, LogR\_new, KNN\_new, SVC\_new, Tree\_new, RF\_new и GB\_new - это объекты, полученные в результате перекрестной проверки (GridSearchCV), примененной к моделям логистической регрессии (Logistic Regression), метода k ближайших соседей (K-Nearest Neighbors), метода опорных векторов (Support Vector Classifier), дерева решений (Decision Tree), случайного леса (Random Forest) и градиентного бустинга (Gradient Boosting), соответственно.



Выполним обучение каждой модели из словаря models\_grid, вычисляем и сохраняем метрики качества и результаты оценки производительности моделей.



## Формирование выводов о качестве построенных моделей на основе выбранных метрик.

Метрики качества модели.

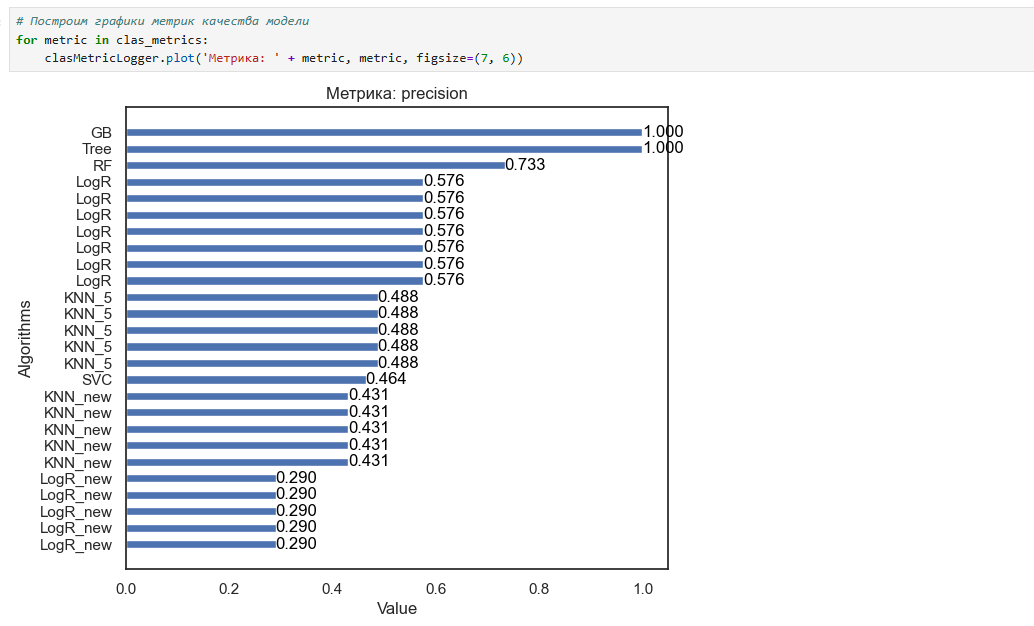
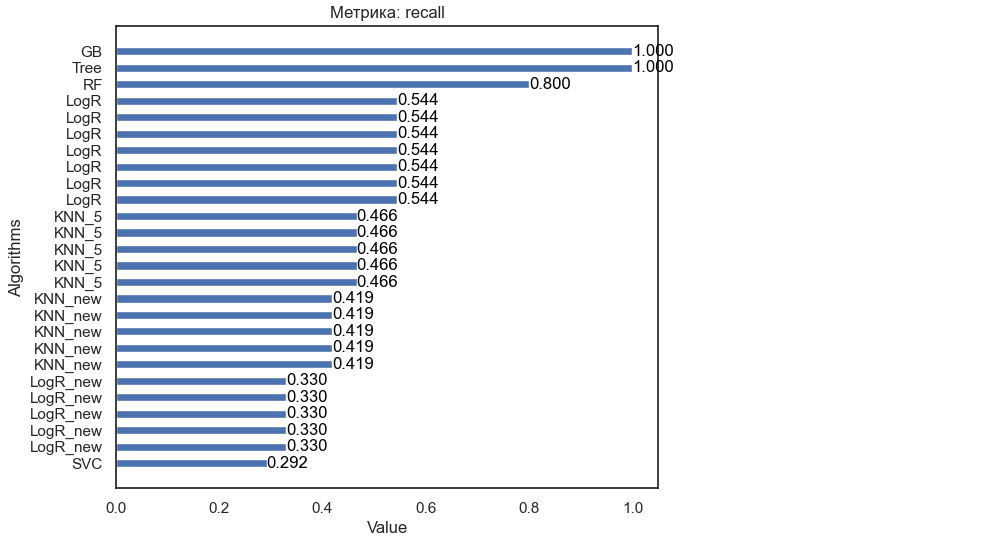
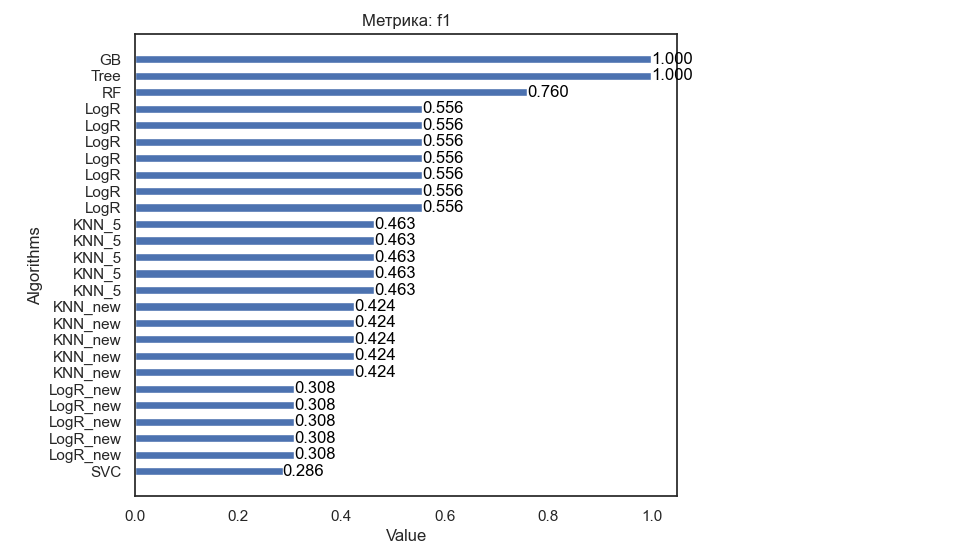






Для наглядности результатов моделей, построим диаграмму.

Построим графики метрик качества модели.

## Вывод

Из предоставленных метрик качества моделей можно сделать следующие выводы:

* Точность (precision): Модели GB, Tree имеют самую высокую точность, равную 1. Модель RF также обладает высокой точностью, около 0.733. Модель LogR имеет точность около 0.57, а модель KNN\_5 - около 0.48. Модели SVC и KNN\_new обладают точностью 0.4. Модель LogR\_new обладает точностью 0.29.
* Полнота (recall): Модели GB, Tree имеют самую высокую точность, равную 1. Модель RF также обладает высокой точностью, около 0.8. Модель LogR имеет точность около 0.54, а модель KNN\_5 - около 0.46. Модель KNN\_new обладает точностью 0.41. Модель LogR\_new обладает точностью 0.33. Модель SVC обладает точностью 0.29.
* F1-мера (f1): Модели GB, Tree имеют самую высокую точность, равную 1. Модель RF также обладает высокой точностью, около 0.76. Модель LogR имеет точность около 0.55, а модель KNN\_5 - около 0.46. Модель KNN\_new обладает точностью 0.42. Модель LogR\_new обладает точностью 0.33. Модель SVC обладает точностью 0.286.

Таким образом, можно сделать вывод, что модели, обученные с использованием логистической регрессии и случайного леса, показывают наилучшее общее качество, выраженное высокими значениями метрик точности, полноты, F1-меры.

## Заключение

В ходе научно-исследовательской работы был проведен разведочный анализ данных набора "Биомеханические особенности пациентов-ортопедов". Мы использовали различные инструменты анализа данных, такие как Pandas, Matplotlib, Seaborn, Numpy и Streamlit на платформе Python с помощью программы Jupyter Notebook.

В процессе работы мы провели анализ структуры данных, заполнили пропущенные значения, выбрали признаки для построения моделей, закодировали категориальные признаки, выполняли масштабирование данных и создали дополнительные признаки для улучшения качества моделей.

Мы также провели корреляционный анализ данных и сформулировали промежуточные выводы о возможности построения моделей машинного обучения.

Для оценки качества моделей мы выбрали метрики, такие как точность, полнота, F1-мера, и рассмотрели более пяти различных моделей, включая как базовые, так и ансамблевые методы.

Мы сформировали обучающую и тестовую выборки, построили базовые решения для выбранных моделей и подобрали оптимальные значения гиперпараметров с использованием методов кросс-валидации.

В итоге мы оценили качество построенных моделей и сравнили их с базовыми моделями, сделав выводы о наилучших подходах к решению задачи классификации или регрессии на основе наших выбранных метрик.

## Список использованных источников информации

1. Абдрахманов, М. И. Devpractice Team. Pandas. Работа с данными. / М. И. Абдрахманов — 2-е изд. — devpractice.ru. 2020. - 170 с.: [Электронный ресурс]. // URL: <https://coderbooks.ru/books/python/pandas_rabota_s_dannymi_abdrahmanov_2020/> (дата обращения: 15.03.2024)
2. Абдрахманов, М. И. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. — devpractice.ru. 2020. - 412 с.: [Электронный ресурс]. // URL: <https://coderbooks.ru/books/python/python_vizualizaciya_dannyh_abdrahmanov_2020/> (дата обращения: 21.03.2024)
3. Методические указания по программному обеспечению «Pandas»: [Электронный ресурс]. // URL: <https://pandas.pydata.org/> (дата обращения: 20.02.2024)
4. Методические указания по программному обеспечению «Seaborn»: [Электронный ресурс]. // URL: <https://seaborn.pydata.org/> (дата обращения: 22.02.2024)
5. Метрики качества. Метрики классификации. Подсчет количества ошибок, доли правильных ответов, точности, полноты. [Электронный ресурс]. // URL: <https://data-scientists.ru/metriki-kachestva-metriki-klassifikacii-podschet-kolichestva-oshibok-doli-pravilnyh-otvetov> (дата обращения: 01.03.2024)
6. Регрессионные модели в Python: [Электронный ресурс]. // URL: <https://nagornyy.me/it/regressionnye-modeli-v-python/> (дата обращения: 11.03.2024)
7. Учебник по машинному обучению: [Электронный ресурс]. // URL: <https://education.yandex.ru/handbook/ml> (дата обращения: 01.03.2024)