ЛАБОРАТОРНАЯ РАБОТА 8. ПОСТРОЕНИЕ ПАЙПЛАЙНА ОДНОМЕРНОЙ РЕГРЕССИИ

**Исследование набора данных**

**Цели и задачи**

Цель лабораторной работы: разработка единого пайплайна для решения задачи регрессии.

Основные задачи:

– реализовать каонвейер для выполнения всех стадий обработки данных при решении задачи одномерной регрессии;

– получение теоретических представлений о задаче регресии;

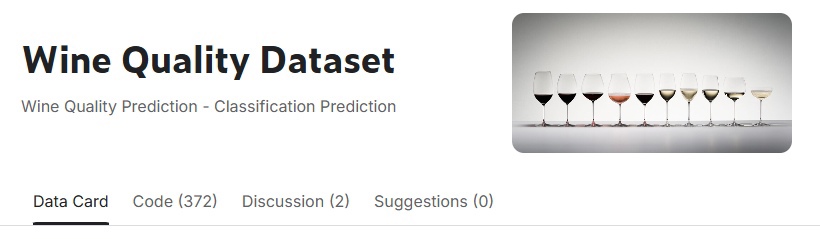
– получение навыков использования пайплайна при решении задачи машинного обучения;

– получение навыков рефакторинга кода в задачах машинного обучения.

**Индивидуальное задание**

**Описание набора данных**

Этот набор данных содержит информацию о различных характеристиках вина. Он включает 12 столбцов, каждый из которых представляет собой определенный атрибут вина. Данные могут использоваться для анализа и прогнозирования качества вина на основе его химических свойств.



Цель набора данных:

Основная цель этого набора данных — исследовать, как различные физико-химические свойства вина влияют на его качество. Это может помочь виноделам и исследователям оптимизировать процесс производства вина и улучшить его качество.

Описание каждого признака и его тип:

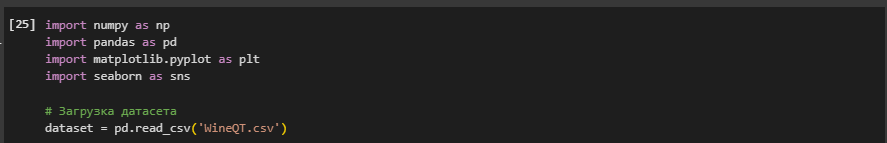
* Id (Тип: Целое число): Уникальный идентификатор.
* fixed acidity (Тип: Число с плавающей точкой): Количество непарных кислот в вине.
* volatile acidity (Тип: Число с плавающей точкой): Количество парных кислот в вине.
* citric acid (Тип: Число с плавающей точкой): Количество лимонной кислоты в вине.
* residual sugar (Тип: Число с плавающей точкой): Количество остаточного сахара в вине после ферментации.
* chlorides (Тип: Число с плавающей точкой): Количество солей (хлоридов) в вине.
* free sulfur dioxide (Тип: Число с плавающей точкой): Количество свободного диоксида серы в вине.
* total sulfur dioxide (Тип: Число с плавающей точкой): Общее количество диоксида серы в вине.
* density (Тип: Число с плавающей точкой): Плотность вина.
* pH (Тип: Число с плавающей точкой): Уровень кислотности вина.
* sulphates (Тип: Число с плавающей точкой): Количество сульфатов в вине.
* alcohol (Тип: Число с плавающей точкой): Количество алкоголя в вине.
* quality (Тип: Целое число): Оценка качества вина (обычно от 0 до 10).

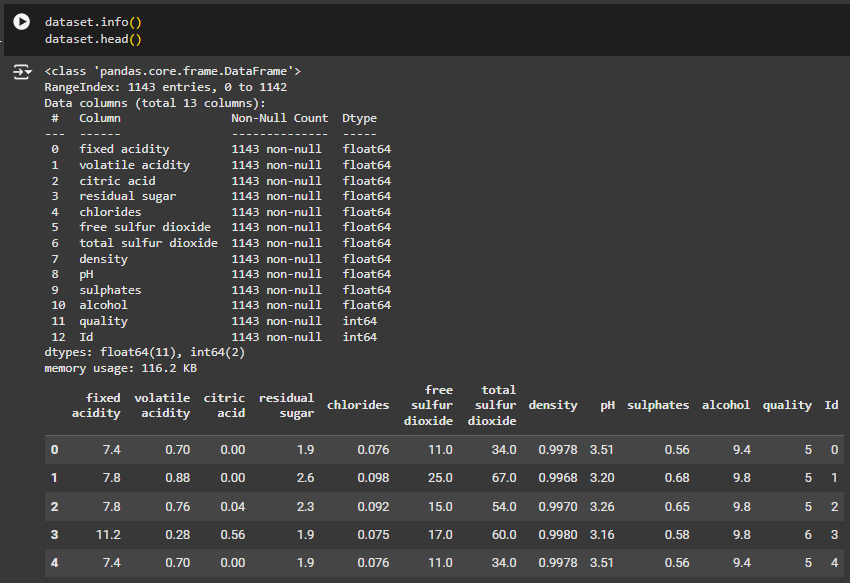
**Форма набора данных**

Характеристики набора данных:

* Количество элементов (строк): 1144
* Количество признаков (столбцов): 13
* Типы данных: Float (11 шт.), int(2 шт.)

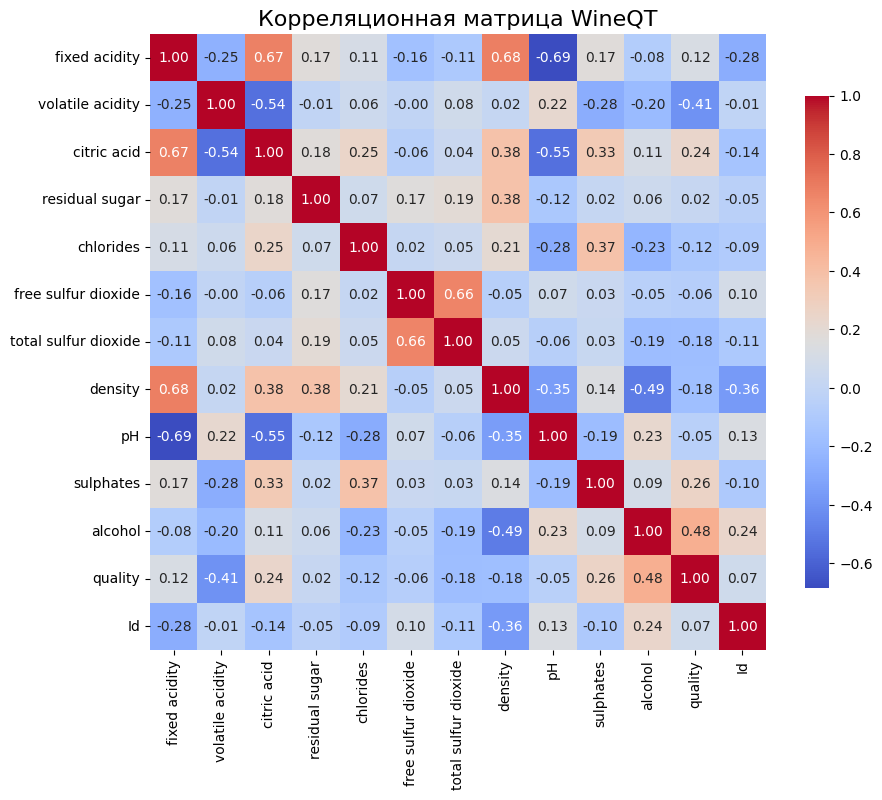
1. Загрузка исходного набора данных.





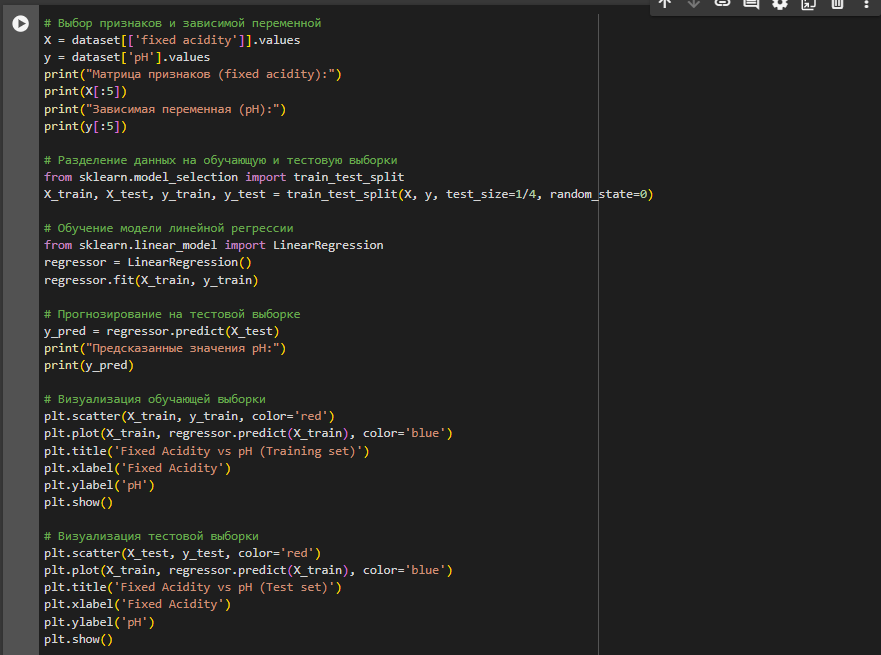
1. После загрузки данных я визуализировал датасет в виде корреляционной матрицы, что бы подобрать наиболее подходящие поля для построения регресии.

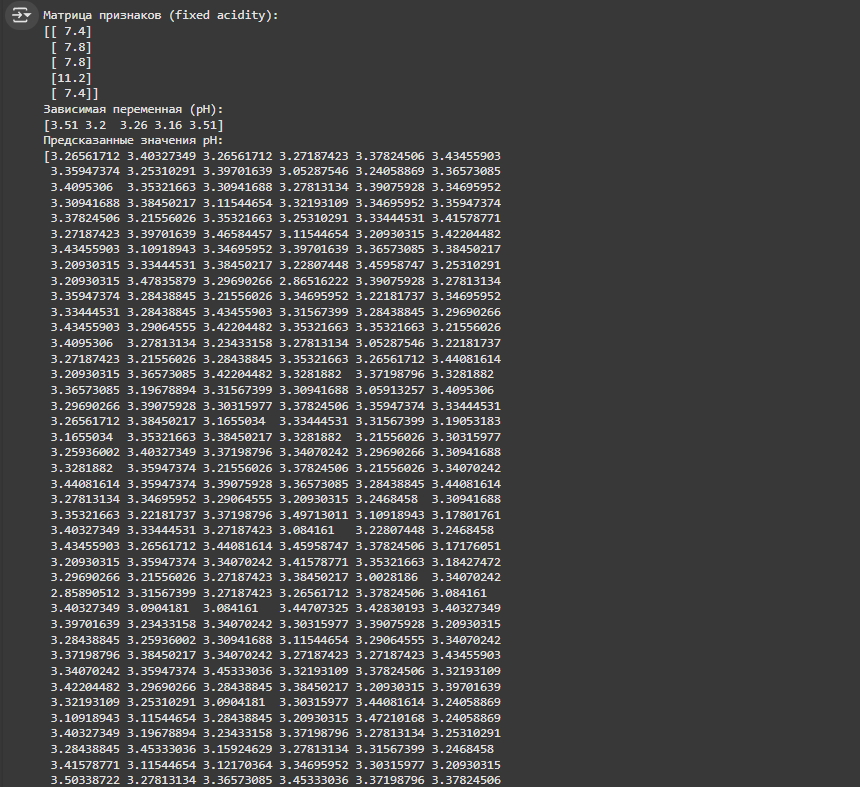


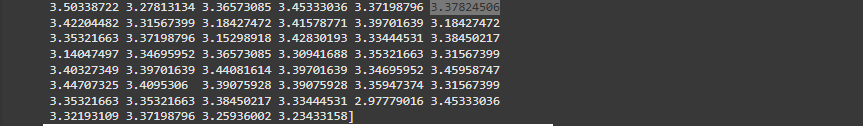


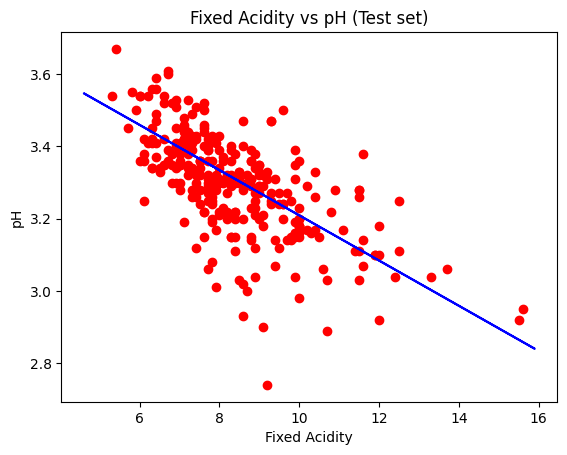
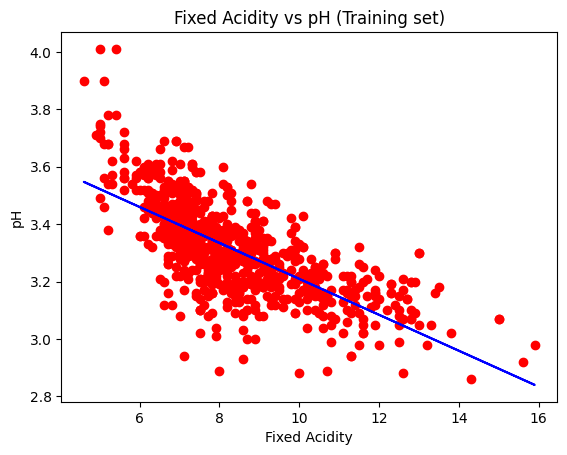
Отсюда замечаем что наиболее подходящие друг к другу поля это fixed acidity и pH, с наименьшим значением (-0.69)

1. После нахождения необходимых полей, строим модель регрессии.







****

**Контрольные вопросы**

1. Почему при реализации линейной модели регрессии нет необходимости выполнять масштабирование признаков?

Масштабирование признаков не обязательно, так как линейная модель может адаптироваться к различным масштабам признаков.

1. Почему при реализации модели линейной регрессии в качестве функции потерь используется квадратичное отклонение, а не модуль отклонения?

Квадратичное отклонение позволяет легче находить оптимальные параметры, так как функция потерь является дифференцируемой, что упрощает оптимизацию.

1. Что именно реализовано в методе fit(X, y) класса LinearRegression?

В методе fit(X, y) происходит обучение модели: вычисляются коэффициенты линейной регрессии на основе входных данных X и целевой переменной y.

1. Что такое p-значение? Как p-значение используется при оптимизации моделей регрессии?

p-значение указывает на вероятность того, что наблюдаемое значение статистики (например, коэффициента) могло бы возникнуть случайно. Используется для проверки значимости признаков в модели регрессии.

1. Поясните назначение метода predict класса LinearRegression.

Метод predict класса LinearRegression используется для предсказания значений целевой переменной на основе новых данных.

1. Поясните назначение метода plot и scatter класса pyplot.

Метод plot используется для создания графиков, а scatter — для построения точечных диаграмм, отображающих взаимосвязь между двумя переменными

1. По какой подвыборке необходимо оценивать точность модели машинного обучения: тестовой или тренировочной?

Точность модели машинного обучения необходимо оценивать на тестовой выборке, чтобы избежать переобучения и получить обобщающую способность модели.